Time Series Analysis: Grad School or Not?

Alison Yao (yy2564)

The Chinese culture has been emphasizing the importance of education for thousands of years. Every Chinese family is proud to send their next generation to prestigious universities and institutions for higher education. And as the job market gets more competitive over time, an undergraduate degree does not seem to suffice sometimes. Therefore, as I almost finish my third year as an undergraduate student, I will be applying for postgraduate programs very soon. It is of interest to me to study and predict the proportion of undergrad graduates who continue to go on to postgraduate study in China.

I collected a sample of 32 observations from China's National Bureau of Statistics, specifically annual data from the education section (https://data.stats.gov.cn/easyquery.htm?cn=C01). In the context of Chinese students, the response is the proportion of undergrad graduates who continue to go on to postgraduate study in China (**Postgrad**); the predictors are the number of regular institutions of higher education per 10000 people (**Institution**), the proportion of full-time teachers in regular institutions of higher education (**Teacher**) and proportion of students studying abroad (**Abroad**). All raw data collected directly from the National Bureau of Statistics are numbers of people. To eliminate size effect, I calculated all predictors and the response in terms of proportion or averaged by 10,000 people:

$$Postgrad\ (\%) = \frac{Number\ of\ New\ Enrollment\ of\ Postgrad\ Students\ (10000\ people)}{Number\ of\ Graduates, Regular\ Graduates\ and\ College\ Students\ (10000\ people)}$$

$$Institution\ (per\ 10000\ people) = \frac{Number\ of\ Regular\ Insitutions\ of\ Higher\ Education}{Number\ of\ Total\ Enrollment\ of\ Postgraduate\ Students\ (10000\ people)}$$

$$Teacher\ (\%) = \frac{Number\ of\ Fulltime\ Teachers\ of\ Regular\ Institutions\ of\ Higher\ Education\ (10000\ people)}{Number\ of\ Teachers\ and\ Staff\ of\ Regular\ Institutions\ of\ Higher\ Education\ (10000\ people)}$$

$$Abroad\ (\%) = \frac{Students\ Studying\ Abroad\ (10000\ people)}{Number\ of\ Graduates, Regular\ Graduates\ and\ College\ Students\ (10000\ people)}$$

The first few observations are listed below:

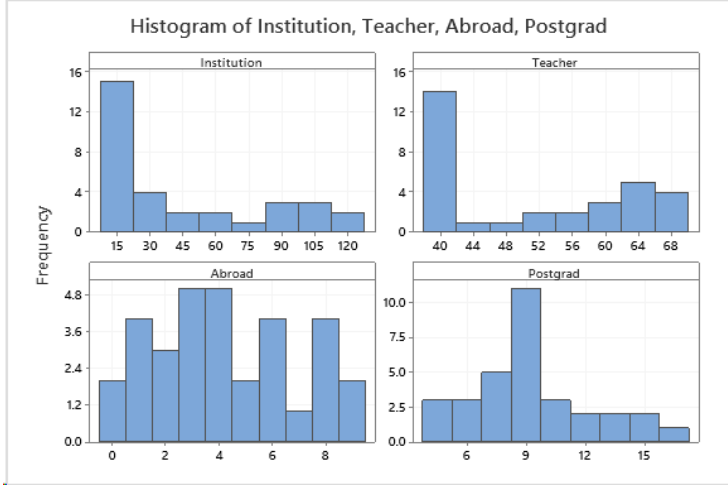
Year	Institution	Teacher	Abroad	Postgrad
1987	88.44256226	40.20618557	0.884022556	7.334022556
1988	95.32169965	39.39393939	0.684629295	6.445750452
1989	106.0795942	40	0.577951389	4.959895833
1990	115.5690297	39.6039604	0.480456026	4.828827362
1991	121.981663	38.61386139	0.472312704	4.833713355

I obtained data from the time period of 1987 to 2018 and not 2019 and 2020 because some recent data are also not recorded (namely Students Studying Abroad, probably because of COVID-19). Let's first look at some basic statistics of the data.

Statistics

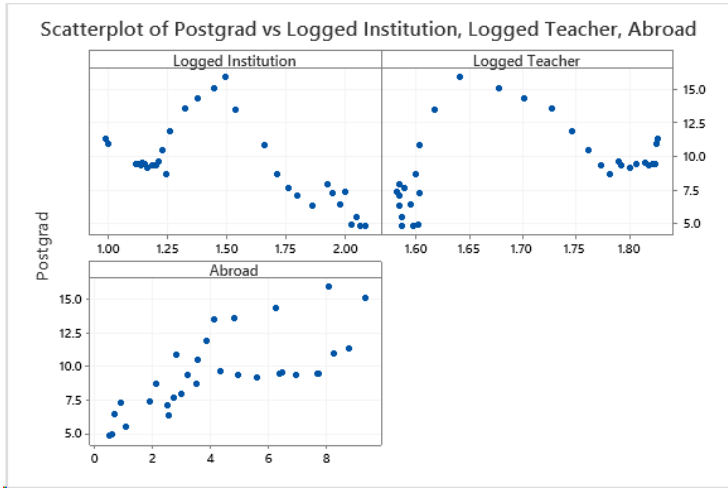
Variable	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum
Institution	32	0	45.49	6.67	37.70	9.75	14.81	25.85	81.43	121.98
Teacher	32	0	50.86	2.05	11.61	38.24	39.45	49.02	62.90	67.07
Abroad	32	0	4.239	0.477	2.699	0.472	2.215	3.706	6.459	9.363
Postgrad	32	0	9.407	0.517	2.927	4.829	7.346	9.388	10.935	15.946

Here are the histograms of the response and predictors.

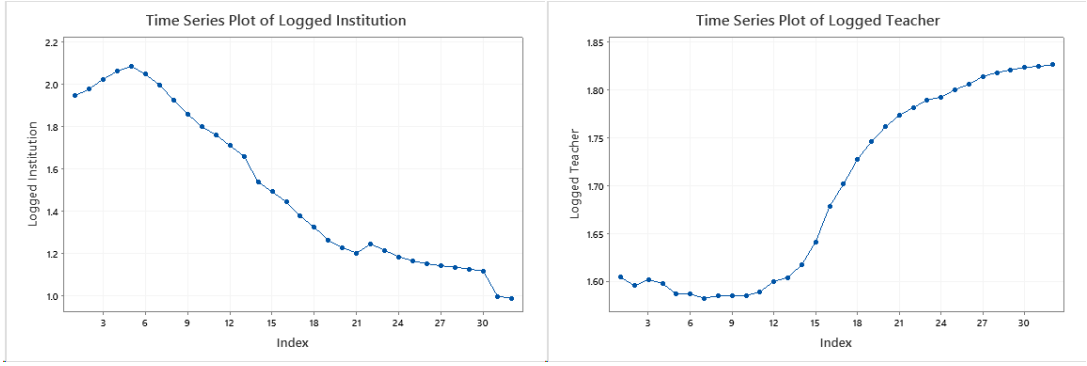


Institution and Teacher have a long right tail, suggesting taking log, while Postgrad has a slight right tail, but it is not too big. Abroad is a little ambiguous. Therefore, I tried several models with and without log. The combination of Logged Institution (base 10), Logged Teacher (base 10), Abroad and Postgrad turns out to fit the assumptions best.

Here is the scatter plot of the response vs all predictors.



There seems to be a linear relationship between Postgrad vs Abroad. The plots of Postgrad vs Logged Institution and Postgrad vs Logged Teacher are going up and then coming down, which is less idea. But these trends are not as apparent in the plots versus time.



Taking time into consideration will most probably mitigate this trend. But first, let us look at the simplest model without time:

Regression Equation

Postgrad = 111.8 - 14.08 Logged Institution - 49.19 Logged Teacher + 0.529 Abroad

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	111.8	18.4	6.09	0.000	
Logged Institution	-14.08	2.70	-5.22	0.000	15.56
Logged Teacher	-49.19	8.47	-5.81	0.000	11.02
Abroad	0.529	0.168	3.16	0.004	3.19

Model Summary

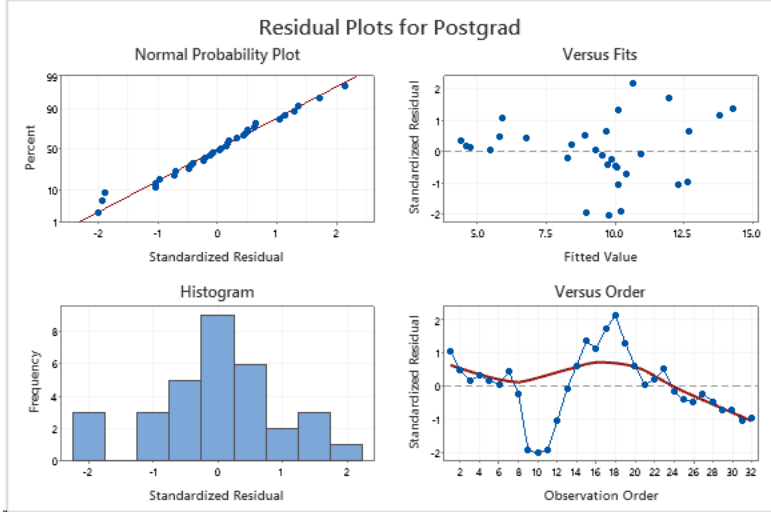
S	R-sq	R-sq(adj)	R-sq(pred)
1.41147	78.99%	76.74%	73.16%

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	3	209.72	69.906	35.09	0.000
Logged Institution	1	54.29	54.289	27.25	0.000
Logged Teacher	1	67.23	67.231	33.75	0.000
Abroad	1	19.84	19.838	9.96	0.004
Error	28	55.78	1.992		
Total	31	265.50			

Durbin-Watson Statistic

Durbin-Watson Statistic = 0.322790

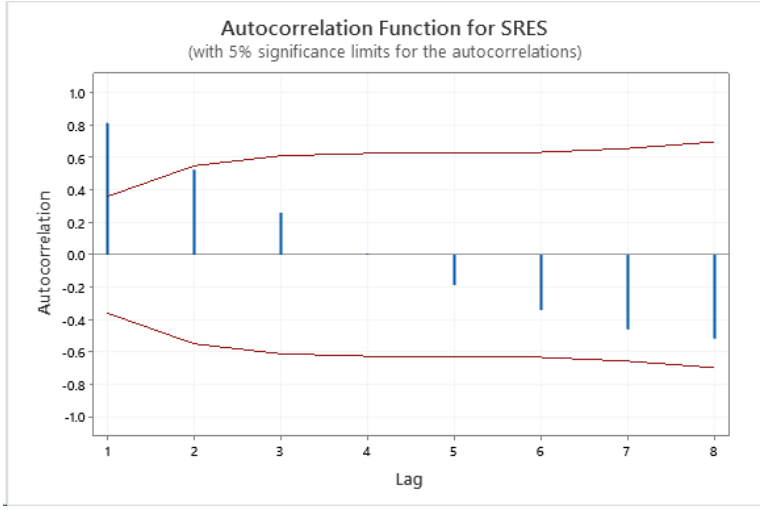


The R-sq is almost 80%, which is pretty good. The F-test and t-tests are highly statistically significant. However, the VIF scores of Logged Institution and Logged Teacher are a little bit high, indicating that there might be redundant variables, which we will investigate further when we do model selection later.

The normal plot is almost a perfect line except for a few unusual points; the standardized residual histogram is not skewed. But unfortunately, the problem of non-constant variance and cyclical effect are quite obvious. There seems to be autocorrelation. Let's check some hypothesis tests to verify that.

The Durbin-Watson value of 0.32 (Λ=3+1=4) is way smaller than QL = 1.039 (1%), indicating autocorrelation. But because of the violation of assumptions, Durbin-Watson is not enough.

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 up to three zeros.

Descriptive Statistics

		Number of Observations	
N	K	≤ K	> K
32	0	15	17

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
4	16.94	0.000

Therefore, we need to attempt to address the autocorrelation of the errors. First, let us try adding Time and Time-sq as new predictors.

Regression Equation

Postgrad = 76.4 - 16.74 Logged Institution - 23.14 Logged Teacher + 0.893 Abroad - 0.282 Time - 0.00613 Time sq

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	76.4	17.3	4.43	0.000	
Logged Institution	-16.74	2.76	-6.08	0.000	37.90
Logged Teacher	-23.14	8.01	-2.89	0.008	23.05
Abroad	0.893	0.125	7.12	0.000	4.16
Time	-0.282	0.133	-2.12	0.044	56.68
Time sq	-0.00613	0.00305	-2.01	0.055	30.56

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.923699	91.64%	90.04%	87.35%

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	5	243.318	48.6636	57.04	0.000
Logged Institution	1	31.491	31.4905	36.91	0.000
Logged Teacher	1	7.113	7.1129	8.34	0.008
Abroad	1	43.235	43.2350	50.67	0.000
Time	1	3.817	3.8174	4.47	0.044
Time sq	1	3.448	3.4484	4.04	0.055
Error	26	22.184	0.8532		
Total	31	265.502			

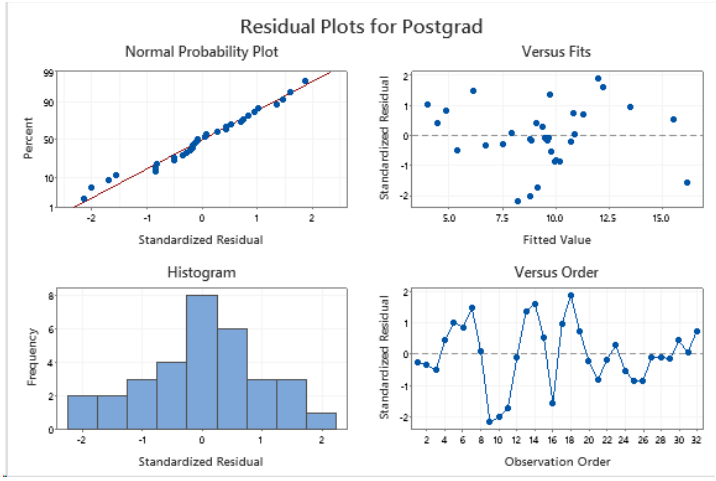
Fits and Diagnostics for Unusual Observations

Obs	Postgrad	Fit	Resid	Std Resid	
9	6.342	8.237	-1.895	-2.16	R
10	7.080	8.811	-1.731	-2.00	R

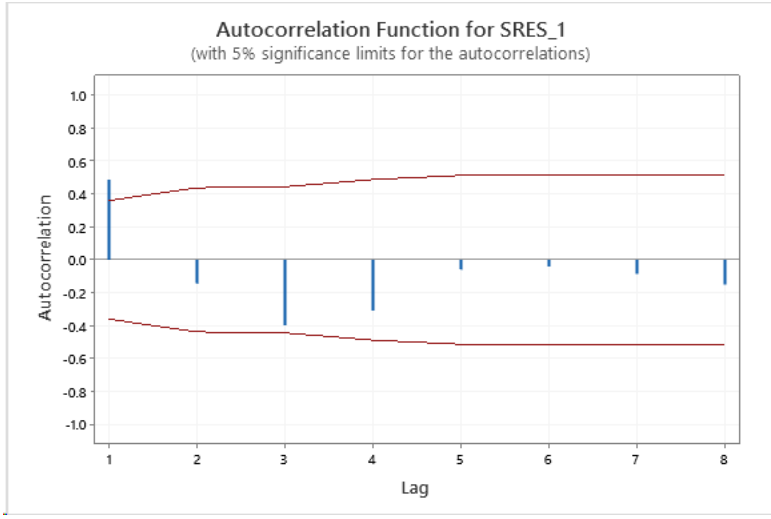
R Large residual

Durbin-Watson Statistic

Durbin-Watson Statistic = 0.938429



The R-sq is now over 90% (increased by over 10%), which is a good sign. Time and Time-sq are not as statistically significant as other predictors, but they are at least marginally significant. Holding everything else fixed, an 1% increase in Institution is associated with an expected decrease of Postgrad by log(1.01) * 16.74 = 0.0723 percentage point; holding everything else fixed, an 1% increase in Teacher is associated with an estimated decrease of 0.100 percentage point in Postgrad. Holding everything else fixed, adding 1 percentage point to Abroad is associated with an expected increase of 0.893 percentage point in Postgrad. Collinearity persists, which will be discussed later. As for assumptions, the normal plot and the std histogram look better, but the non-constant variance continues to exist and there still seems to be a cyclical effect. The Durbin-Watson test falls between QL=0.917 and QU=1.597 (1%) and its assumptions do not hold well, so the result is ambiguous. Let us look at further tests. The ACF plot is as follows:



The autocorrelation is slightly lessened, but it still exists when Lag=1. The runs test suggests the same, with a slightly larger p-value of 0.012, still showing evidence to reject the null hypothesis that there is no autocorrelation.

Descriptive Statistics

Number of Observations			
N	K	≤ K	> K
32	0	17	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

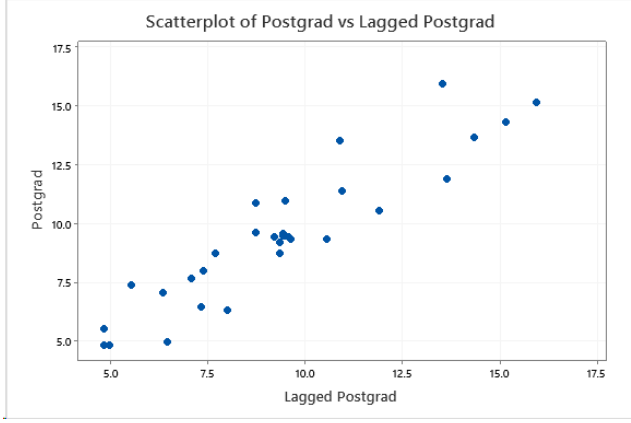
Number of Runs		
Observed	Expected	P-Value
10	16.94	0.012

The collinearity problem stated above leads us to perform a best subset model selection before moving on to further address the autocorrelation problem.

Vars	R-Sq	R-Sq (adj)	PRESS	R-Sq (pred)	Mallows Cp	S	AICc	BIC	Cond No	Logged Institution	Logged Teacher	Abroad	Time	Time sq
1	53.7	52.1	139.7	47.4	116.2	2.0250	140.760	144.300	1.000			X		
1	35.3	33.2	188.0	29.2	173.3	2.3925	151.433	154.973	1.000	X				
2	71.5	69.6	90.9	65.8	62.6	1.6148	127.814	132.195	38.723	X	X			
2	68.9	66.7	100.3	62.2	70.9	1.6884	130.667	135.049	8.625			X		X
3	88.4	87.2	43.0	83.8	12.1	1.0482	101.858	106.879	30.588	X		X		X
3	80.8	78.7	66.2	75.1	35.8	1.3501	118.060	123.081	89.677	X		X	X	
4	90.3	88.9	35.5	86.6	8.0	0.97434	99.072	104.506	118.393	X	X	X	X	
4	90.2	88.8	38.0	85.7	8.5	0.98133	99.529	104.963	136.575	X	X	X		X
5	91.6	90.0	33.6	87.4	6.0	0.92370	97.755	103.348	513.919	X	X	X	X	X

Interestingly, the current model with 5 predictors is the best, with the highest adjusted R-sq and predictive R-sq and lowest Mallows' Cp (also Cp=5+1=6) and AICc. The 4-predictor model that exclude Time or Time-sq comes close, but the Cp score goes up by 2-2.5. Therefore, I am not going to discard any predictor at this point.

To resume dealing with autocorrelation, we lag the response by 1 and use the Lagged Postgrad as a predictor. Just for intuition, we can see this scatter plot:



Lagging the response seems sensible because Lagged Postgrad and Postgrad have a linear relationship. The output of the regression is:

Method

Rows unused 1

Regression Equation

Postgrad = 56.6 - 7.72 Logged Institution - 24.41 Logged Teacher + 0.421 Abroad - 0.003 Time - 0.00229 Time sq + 0.515 Lagged Postgrad

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	56.6	17.8	3.19	0.004	
Logged Institution	-7.72	4.37	-1.77	0.090	112.15
Logged Teacher	-24.41	7.25	-3.36	0.003	22.68
Abroad	0.421	0.201	2.09	0.047	12.56
Time	-0.003	0.193	-0.02	0.987	133.63
Time sq	-0.00229	0.00343	-0.67	0.511	46.05
Lagged Postgrad	0.515	0.182	2.83	0.009	12.50

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.831113	93.65%	92.06%	89.21%

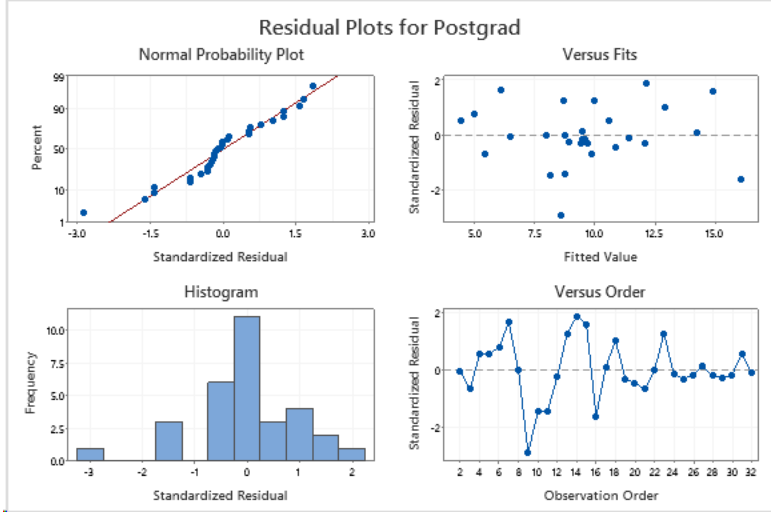
Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	6	244.486	40.7477	58.99	0.000
Logged Institution	1	2.159	2.1587	3.13	0.090
Logged Teacher	1	7.821	7.8206	11.32	0.003
Abroad	1	3.019	3.0191	4.37	0.047
Time	1	0.000	0.0002	0.00	0.987
Time sq	1	0.307	0.3072	0.44	0.511
Lagged Postgrad	1	5.541	5.5407	8.02	0.009
Error	24	16.578	0.6907		
Total	30	261.064			

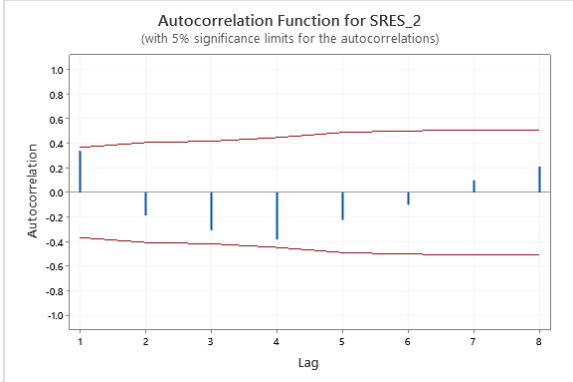
Fits and Diagnostics for Unusual Observations

Obs	Postgrad	Fit	Resid	Std Resid	
9	6.342	8.587	-2.245	-2.88	R

R Large residual



The cyclical effect seems to be mitigated and the con-constant variance problem also seems better. From the ACF plot and the runs test, we can see that autocorrelation is resolved.



Descriptive Statistics

Number of Observations			
N	K	≤ K	> K
31	0	19	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
13	15.71	0.296

However, from the results of the t-tests, we can see that Time and Time-sq are not statistically significant anymore. This leads us to do another best subset model selection:

31 cases used, 1 cases contain missing values

Vars	R-Sq	R-Sq (adj)	PRESS	R-Sq (pred)	Mallows Cp	S	AICc	BIC	Cond No	Logged Institution	Logged Teacher	Abroad	Time	Time-sq	Lagged Postgrad
1	85.0	84.5	44.6	82.9	29.7	1.1622	102.115	105.528	1.000						X
1	53.0	51.4	139.6	46.5	150.5	2.0564	137.495	140.908	1.000			X			
2	86.9	85.9	40.5	84.5	24.6	1.1067	100.641	104.838	5.276			X			X
2	85.3	84.2	47.2	81.9	30.6	1.1710	104.147	108.345	2.609		X				X
3	92.3	91.4	26.3	89.9	6.1	0.86259	86.927	91.697	47.369		X		X		X
3	91.1	90.2	30.1	88.5	10.5	0.92576	91.309	96.079	82.475	X	X				X
4	93.5	92.5	24.7	90.6	3.6	0.80810	84.811	89.915	112.941	X	X	X			X
4	92.5	91.4	28.9	88.9	7.2	0.86540	89.058	94.162	109.379		X	X	X		X
5	93.6	92.4	26.7	89.8	5.0	0.81433	87.441	92.609	243.512	X	X	X		X	X
5	93.5	92.2	27.0	89.7	5.4	0.82183	88.010	93.178	873.386	X	X	X	X		X
6	93.6	92.1	28.2	89.2	7.0	0.83111	91.116	96.043	1203.063	X	X	X	X	X	X

This time, the 4-predictor semi log model is the best, with Adjusted R-sq, predictive R-sq Mellows' Cp and AICc all pointing to selecting Logged Institution, Logged Teacher, Abroad and Lagged Postgrad. Time and Time-sq no longer add predictive power after adding lagged response into the equation. It does make sense that this year's proportion of undergrad graduates who continue to go on to postgraduate study in China is related to last year's proportion. Let's fit this 4-predictor regression model:

Method

Rows unused 1

Regression Equation

Postgrad = 59.2 - 6.88 Logged Institution - 27.37 Logged Teacher + 0.310 Abroad + 0.6108 Lagged Postgrad

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	59.2	12.8	4.61	0.000	
Logged Institution	-6.88	1.85	-3.71	0.001	21.37
Logged Teacher	-27.37	5.75	-4.76	0.000	15.05
Abroad	0.310	0.101	3.07	0.005	3.34
Lagged Postgrad	0.6108	0.0816	7.49	0.000	2.67

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.808096	93.50%	92.50%	90.56%

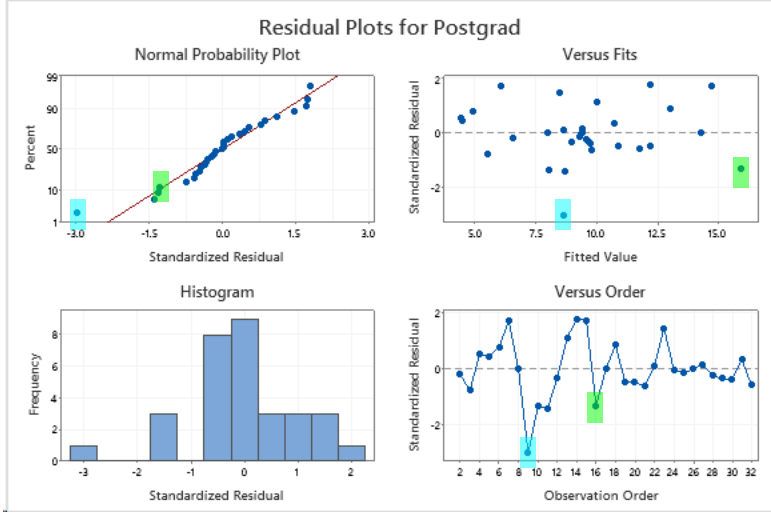
Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	4	244.086	61.0215	93.45	0.000
Logged Institution	1	8.994	8.9938	13.77	0.001
Logged Teacher	1	14.826	14.8256	22.70	0.000
Abroad	1	6.161	6.1615	9.44	0.005
Lagged Postgrad	1	36.587	36.5874	56.03	0.000
Error	26	16.978	0.6530		
Total	30	261.064			

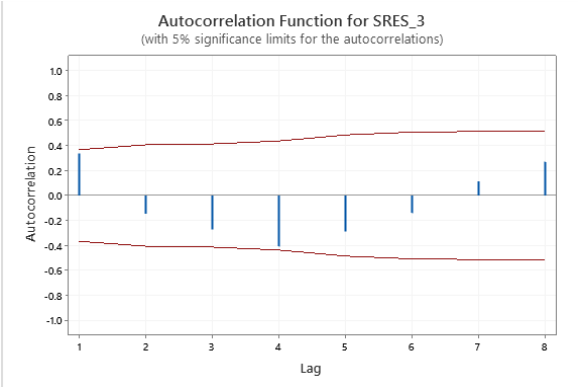
Fits and Diagnostics for Unusual Observations

Obs	Postgrad	Fit	Resid	Std Resid	
9	6.342	8.654	-2.312	-2.98	R

R Large residual



The R-sq is very high, meaning that the regression accounts for more than 90% of variability. The F-statistics and t-statistics are all highly statistically significant. Holding the others fixed, adding 1 percentage point to last year's Postgrad proportion is associated with an expected increase of 0.6108 percentage point in the current year's Postgrad proportion. There is still a collinearity problem, but best subset and test statistics show that all of these predictors are necessary. The slope of the lagged response is not close to 1 so there is no suggestion that differencing the response might help. Let's double-check that there is no autocorrelation problem anymore:



Descriptive Statistics

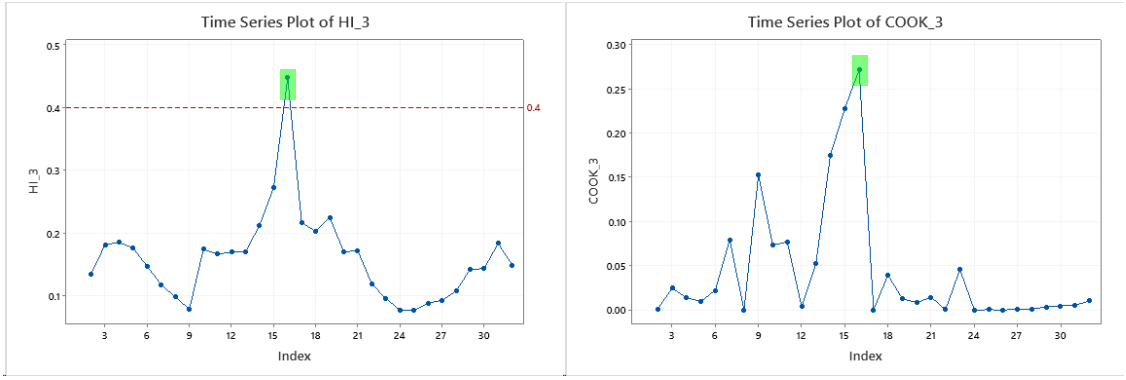
Number of Observations			
N	K	≤ K	> K
31	0	16	15

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
13	16.48	0.203

Now, let's look at the unusual observations. From the four-in-one plot, it is quite obvious that Year 1995 (highlighted in blue) is an outlier. The standardized residual of Year 1995 is -2.95, which is especially low. From the diagnostics plots below, we can also see that Year 2002 (highlighted in green) is a leverage point. Year 2002 has a high leverage value. Although all Cook values are much smaller than 1, Year 2002 does appear unusual. For now, I am not going to look at other points that can potentially be unusual based on Cook_3 values.



What happened in Year 1995 and Year 2002? In 1995, the number of new enrolled postgraduate students (10000 people) is increasing at a steady rate, but the number of graduates (undergraduates and college students) (10000 people) suddenly rose, causing the proportion (Postgrad) to drop when the denominator is too big. The reason behind it is that the Chinese government permitted private institutions of higher education in the early 1980s, therefore created much more undergraduates from 1995 to 2005 (increasing steadily after 1995, but there is a jump from 1994 to 1995). In 2002, it was the opposite: the numerator of Students Studying Abroad after graduating from undergraduate universities and colleges, making Abroad unusually big. There might be a lot of reasons why this happened. I believe it was because 2002 was the end of Chairman Deng's era, during which the Chinese Economic Reform took place and Chinese economy peaked at the end of the era. The income level of most Chinese families increased and more could afford the expenditure of sending their children abroad for better education (most people going abroad are not funded by the state).

Now, to address these unusual points, let's first deal with the outlier Year 1995. Since there seems to be only one clear outlier so far, I imputed the predictor of Year 1995 to be 7.873 (6.342 is a little too low) and used the adjusted Postgrad as the new response. Here is the result of best subset model selection:

31 cases used, 1 cases contain missing values

Vars	R-Sq	R-Sq (adj)	PRESS	R-Sq (pred)	Mallows Cp	S	AICc	BIC	Cond No	Logged Institution	Logged Teacher	Abroad	Time	Time sq	Lagged Postgrad
1	85.9	85.5	40.9	83.9	45.4	1.1094	99.234	102.647	1.000						X
1	52.8	51.2	136.7	46.2	216.0	2.0322	136.760	140.173	1.000			X			
2	87.6	86.7	37.4	85.3	38.7	1.0590	97.909	102.107	5.276			X			X
2	86.5	85.5	42.5	83.3	44.7	1.1075	100.686	104.883	2.609		X				X
3	94.3	93.7	19.4	92.4	6.3	0.73194	76.744	81.514	47.369		X		X		X
3	92.7	91.9	24.5	90.4	14.4	0.82617	84.252	89.022	82.475	X	X				X
4	95.2	94.5	18.9	92.6	3.6	0.68284	74.369	79.473	112.941	X	X	X			X
4	94.5	93.6	22.0	91.3	7.4	0.73362	78.816	83.920	109.379		X	X	X		X
5	95.3	94.4	21.0	91.7	5.1	0.68970	77.143	82.311	243.512	X	X	X		X	X
5	95.2	94.3	20.6	91.9	5.6	0.69636	77.739	82.907	873.386	X	X	X	X		X
6	95.3	94.2	21.9	91.4	7.0	0.70208	80.656	85.583	1203.063	X	X	X	X	X	X

The best model uses the same predictors as before, but the response has been adjusted. Since we manually adjusted one predictor value, the model is not applicable to Year 1995 or any other year if we were to have any. Here is the Minitab output:

Method

Rows unused 1

Regression Equation

Adj Postgrad = 61.4 - 7.04 Logged Institution - 28.51 Logged Teacher + 0.3137 Abroad + 0.6053 Lagged Postgrad

Coefficients

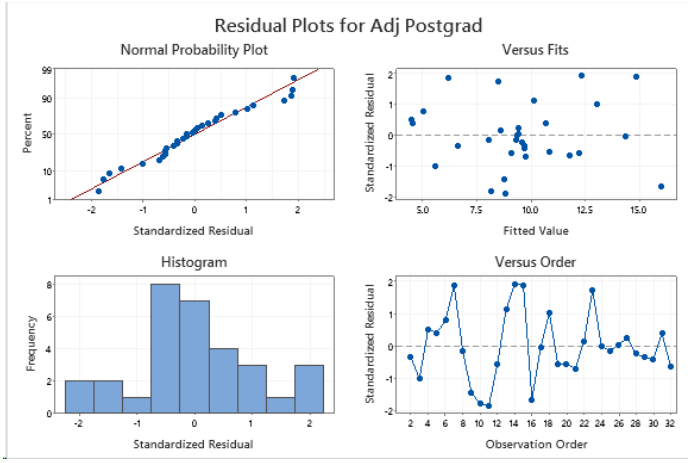
Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	61.4	10.9	5.66	0.000	
Logged Institution	-7.04	1.57	-4.50	0.000	21.37
Logged Teacher	-28.51	4.85	-5.87	0.000	15.05
Abroad	0.3137	0.0853	3.68	0.001	3.34
Lagged Postgrad	0.6053	0.0689	8.78	0.000	2.67

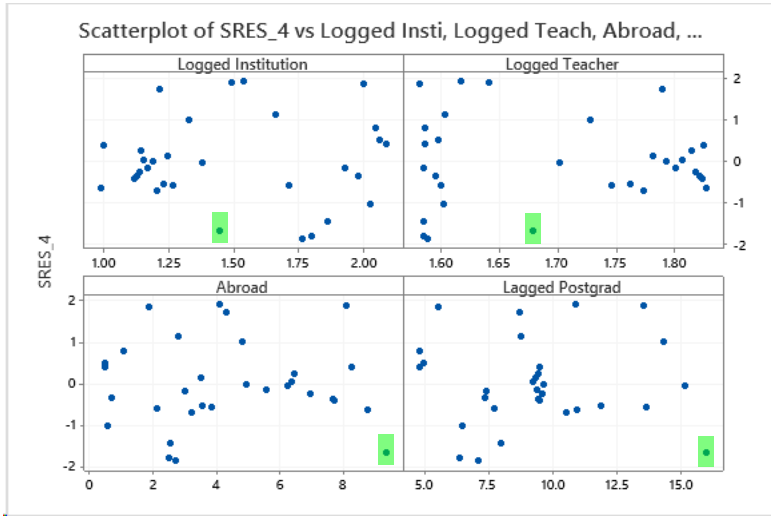
Model Summary

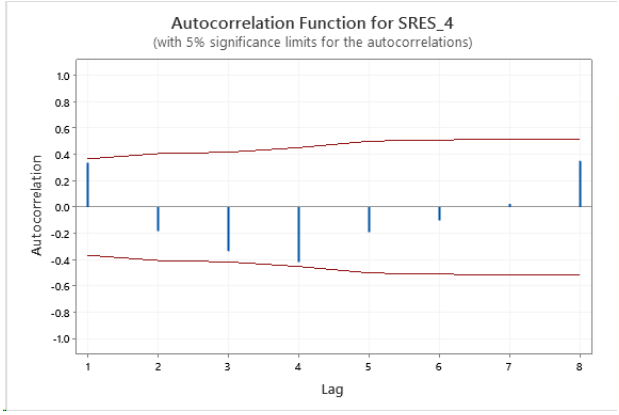
S	R-sq	R-sq(adj)	R-sq(pred)
0.682842	95.22%	94.49%	92.56%

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	4	241.738	60.4345	129.61	0.000
Logged Institution	1	9.422	9.4222	20.21	0.000
Logged Teacher	1	16.079	16.0791	34.48	0.000
Abroad	1	6.306	6.3059	13.52	0.001
Lagged Postgrad	1	35.938	35.9381	77.08	0.000
Error	26	12.123	0.4663		
Total	30	253.861			







Descriptive Statistics

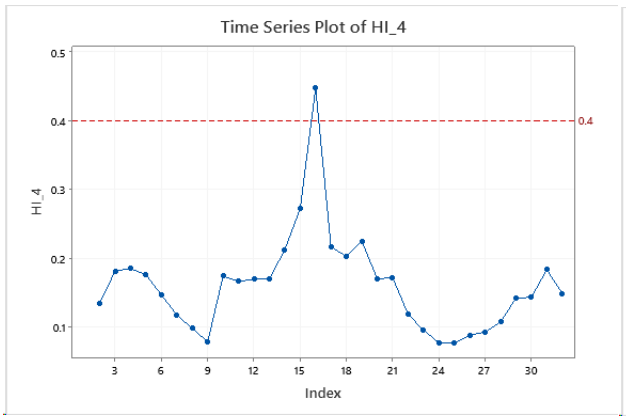
	Number of Observations		
N	K	≤ K	> K
31	0	17	14

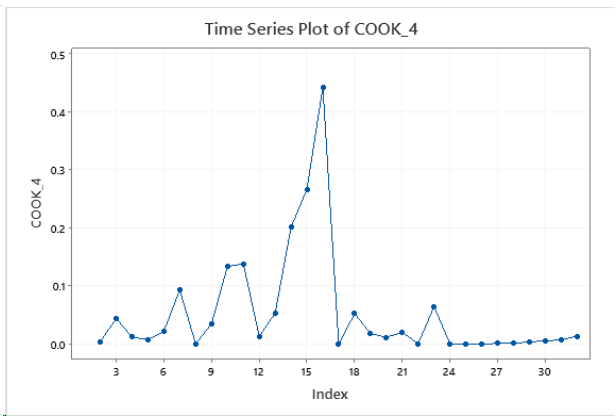
Test

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	16.35	0.216

Some non-constant variance problem persists, especially when looking at the std versus each residual plot. There is still some collinearity, but it does not seem to be a big problem. Autocorrelation has been resolved. Not surprisingly, Year 2002 continues to be a leverage point according to the following plots:





I also tried adding an indicator variable Year1995 to deal with the outlier. The R-sq, Adj R-sq coefficients, F-statistics, t-statistics and the four-in-one plot are all very similar to the imputed version. Notably, the t-test for Year1995 yields a p-value of 0.001, showing that Year 1995 is indeed an outlier. Collinearity is still there and the diagnostics still shows strong evidence that Year 2002 is a leverage point.

Now, let's continue with the imputed version and deal with the leverage point by adding an indicator variable Year2002. The best subset model selection is shown below.

31 cases used, 1 cases contain missing values

Total Vars	R-Sq	R-Sq (adj)	PRESS		R-Sq (pred)		Mallows Cp	S	AICc	BIC
2	86.0	85.0	*			*	56.9	1.1259	101.710	105.908
2	54.1	50.8	*			*	243.7	2.0398	138.555	142.753
3	87.8	86.4	*			*	48.4	1.0710	100.346	105.116
3	86.7	85.2	*			*	54.8	1.1174	102.973	107.743
4	94.5	93.6	*			*	11.5	0.73584	79.003	84.107
4	93.4	92.4	*			*	17.4	0.79994	84.182	89.286
5	95.7	94.9	*			*	6.0	0.65874	74.295	79.464
5	95.0	94.0	*			*	10.1	0.71099	79.028	84.196
6	96.1	95.1	*			*	6.1	0.64571	75.466	80.393
6	95.7	94.7	*			*	7.9	0.67074	77.825	82.751
7	96.1	94.9	*			*	8.0	0.65844	79.384	83.718

Total Vars	Cond No	Logged Institution	Teacher	Logged	Abroad	Time	Time sq	Lagged Postgrad
2	2.419							X
2	2.069				X			
3	6.304				X			X
3	4.276	X						X
4	47.767	X				X		X
4	13.156	X			X			X
5	130.248	X	X		X			X
5	132.169	X	X		X	X		X
6	272.553	X	X		X		X	X
6	888.578	X	X		X	X		X
7	1206.723	X	X		X	X	X	X

At your request, the best subsets procedure included these variables in every model: Year2002

The 5-predictor model and the 6-predictor model are both very good. For simplicity, I will choose the one with fewer predictors. The chosen 5-predictor model adds Year2002 to the previous model. Again, the model does not apply to Year 1995 because we imputed the predictor. Let's look at some details:

Method

Rows unused 1

Regression Equation

Adj Postgrad = 56.4 - 5.97 Logged Institution - 26.95 Logged Teacher + 0.3866 Abroad + 0.6578 Lagged Postgrad - 1.518 Year2002

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	56.4	10.9	5.19	0.000	
Logged Institution	-5.97	1.64	-3.65	0.001	25.02
Logged Teacher	-26.95	4.77	-5.65	0.000	15.62
Abroad	0.3866	0.0926	4.17	0.000	4.24
Lagged Postgrad	0.6578	0.0732	8.98	0.000	3.23
Year2002	-1.518	0.886	-1.71	0.099	1.75

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.658744	95.73%	94.87%	*

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	5	243.012	48.6025	112.00	0.000
Logged Institution	1	5.789	5.7890	13.34	0.001
Logged Teacher	1	13.840	13.8404	31.89	0.000
Abroad	1	7.558	7.5580	17.42	0.000
Lagged Postgrad	1	35.016	35.0162	80.69	0.000
Year2002	1	1.275	1.2745	2.94	0.099
Error	25	10.849	0.4339		
Total	30	253.861			

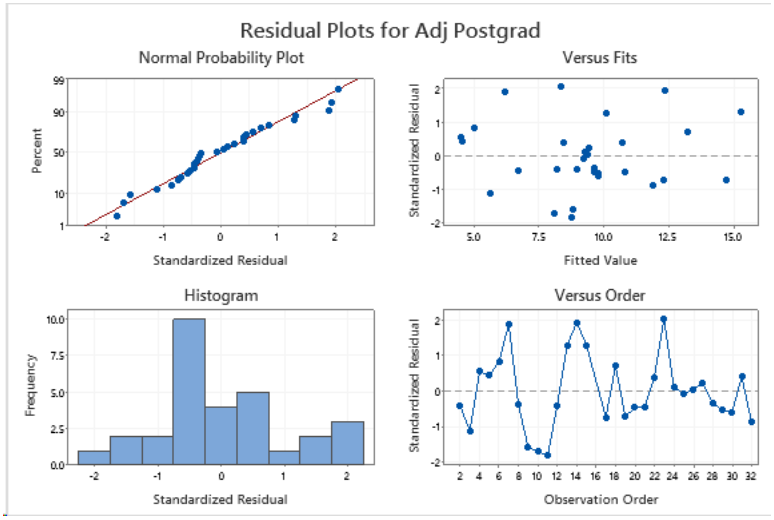
Fits and Diagnostics for Unusual Observations

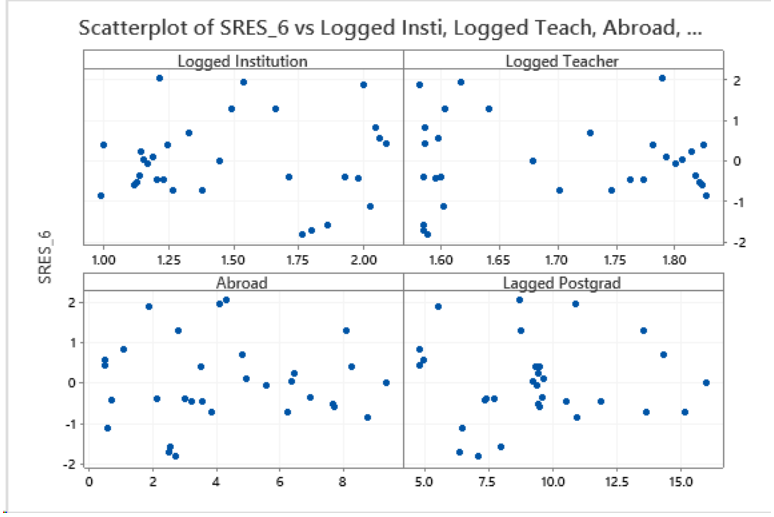
Obs	Adj Postgrad	Fit	Resid	Std Resid	
16	15.154	15.154	0.000	*	X
23	9.621	8.346	1.274	2.05	R

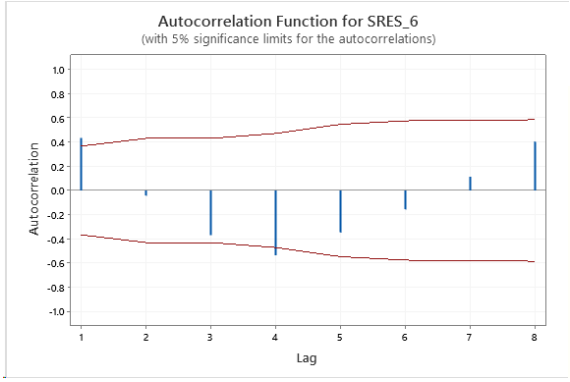
R Large residual
X Unusual X

The R-sq and adjusted R-sq have improved to about 95%, meaning that our predictors account for around 95% of variability. The F-test and t-tests are all highly statistically significant, except for Year2002, which is marginally significant (it makes sense because it is a leverage point). Holding everything else fixed, adding 1% to Institution is associated with an expected decrease of Postgrad by 0.026 percentage point; holding the others fixed, adding 1% to Teacher is associated with an expected decrease of 0.116 percentage point in Postgrad. Holding everything else fixed, adding 1 percentage point to Abroad is associated with an expected increase of 0.3866 percentage point in Postgrad. Holding the others fixed, adding 1 percentage point to last year's Postgrad proportion is associated with an expected increase of 0.6578 percentage point in the current year's Postgrad proportion. The coefficient of Year2002 shows

that holding other variable fixed, Postgrad proportion in Year 2002 is 1.518 percentage point lower than expected.







Descriptive Statistics

Number of Observations

N	K	≤ K	> K
31	0	17	14

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
13	16.35	0.216

The assumption violation and the collinearity situation have not changed much. The std histogram is starting to show a little bit of right tail though. The ACF plot shows that some mild

autocorrelation reoccurs, but the runs test stays exactly the same as the last one, showing no evidence of autocorrelation.

In conclusion, the best model for these time series data stabilizes with Logged Institution, Logged Teacher, Abroad and Lagged Postgrad if we do not deal with the unusual points, so these predictors prove to be statistically significant and possessing predictive power. Autocorrelation is well-addressed after using lagging. After dealing with two unusual points, we add one more indicator variable to the model. There might be a hint of evidence that autocorrelation came back after taking addressing the leverage point, so in the future, we might improve by using more sophisticated tools of time series analysis to look deeper into this question. For the current best model, I actually find it counterintuitive because I was expecting things like a higher Teacher proportion contributing to a higher Postgrad proportion given everything else. But I suppose people's attitude toward higher education can be influenced by government policy and cultural influences. And the bar of going to universities and colleges and getting higher degrees is arguably getting lower these days. A multiple regression model like this might not be able to capture such complicated factors.