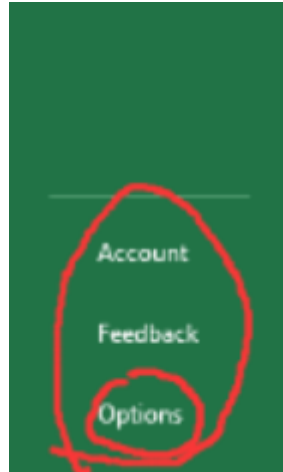


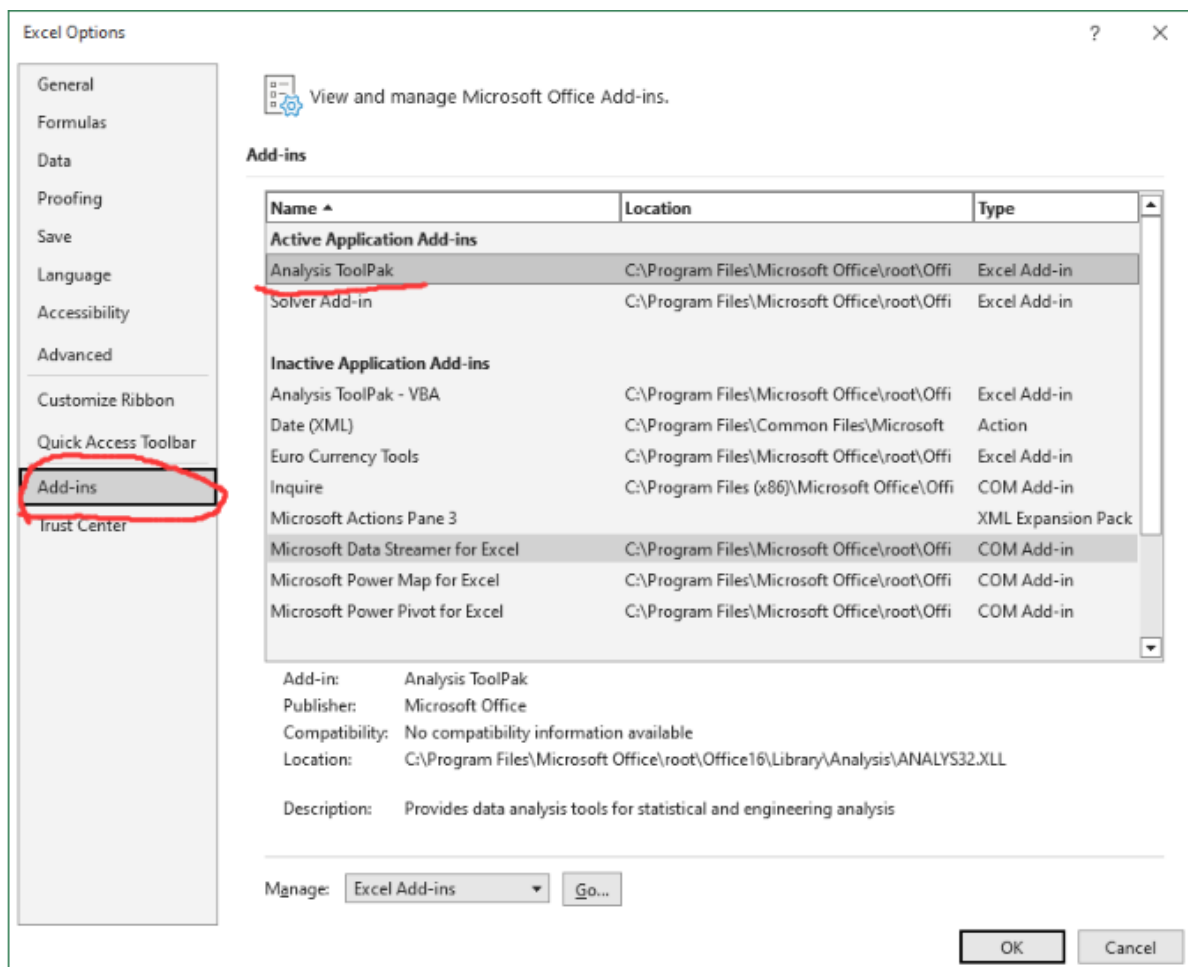
# Lake Huron

## Excel

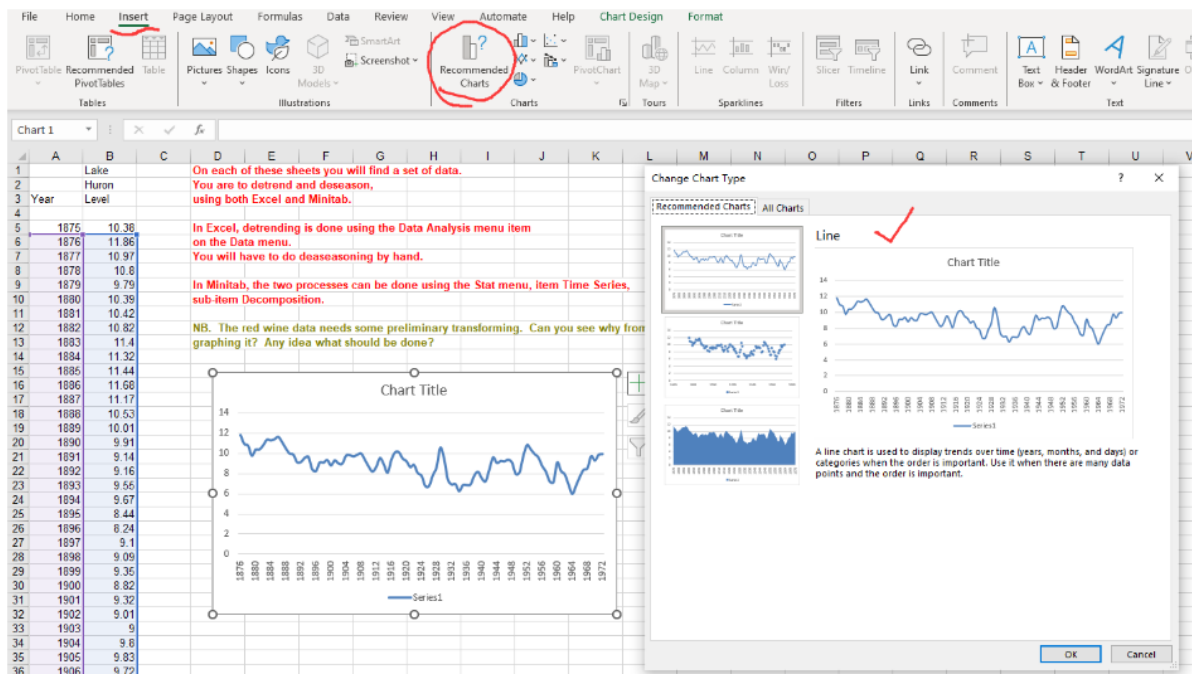
Open Excel, Click File and Find the Option:



Then Add the Analysis ToolPak:



Select all data but not 1875, and the click Inter Button and click recommendation charts button:



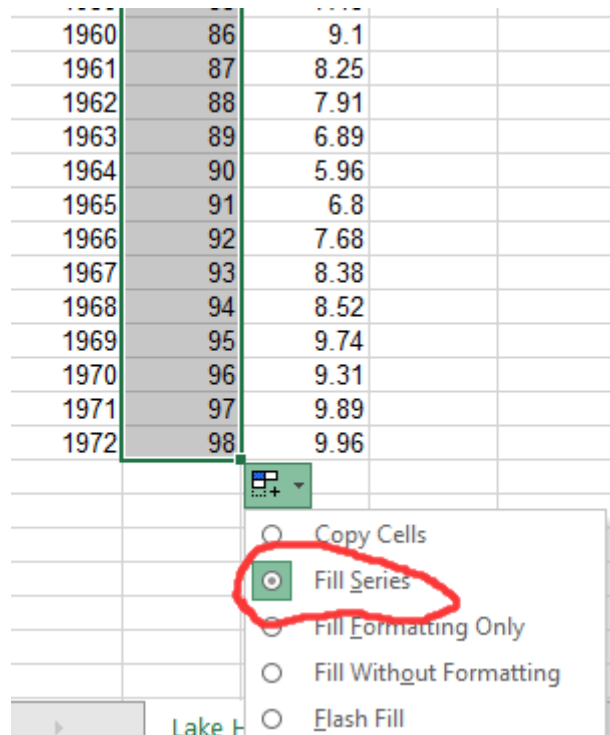
A Line Chart is drawn as shown above, date is the x-value and level is the y value.

Just from the changes in the polyline in the above figure, we can generally believe that there is a downward trend over time.

Next, we proceed to regression analysis:

Year	Series	Lake Huron Level
1875	1	10.38
1876		11.86
1877		10.97
1878		10.8
1879		9.79
1880		10.39
1881		10.42
1882		10.82
1883		11.4
1884		11.32
1885		11.44
1886		11.68
1887		11.17
1888		10.53
1889		10.01
1890		9.91
1891		9.14
1892		9.16
1893		9.55
1894		9.67
1895		8.44
1896		8.24
1897		9.1
1898		9.09

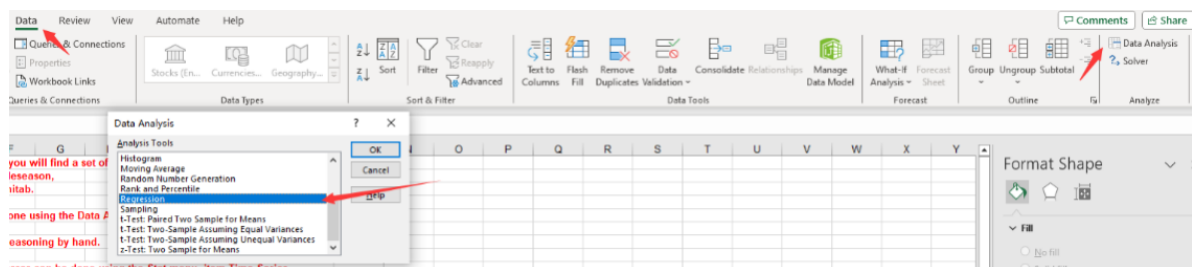
Hold on and move is to the end of the data column, and then select fill series:



When performing regression analysis on time series data, why not use time as the x value?

Time is not used as an independent variable because time usually does not provide explanation or causality about the dependent variable, which may introduce multicollinearity problems, increase model complexity, and reduce model generalization ability. We focus more on capturing the intrinsic structure of time series, such as trends and seasonality, to improve forecasting performance.

Then we can start doing regression analysis:



After Clicking OK:

The image shows an Excel spreadsheet with columns A, B, and C. Column A contains years from 1875 to 1899. Column B contains index values from 1 to 25. Column C contains data levels from 10.38 to 9.35. A red arrow points from the 'Input Y Range' field in the 'Regression' dialog box to the data in column C (rows 5 to 25). Another red arrow points from the 'Input X Range' field to the index values in column B (rows 5 to 25). A third red arrow points from the 'Residuals' checkbox to the 'Residuals' button. The 'Regression' dialog box is open, showing the following settings:

- Input Y Range:** \$C\$5:\$C\$102
- Input X Range:** \$B\$5:\$B\$102
- Labels:** ☐
- Constant is Zero:** ☐
- Confidence Level:** 95 %
- Output options:**
  - ☐ Output Range:
  - ☒ New Worksheet Ply:
  - ☐ New Workbook
- Residuals:**
  - ☒ Residuals
  - ☐ Standardized Residuals
  - ☐ Residual Plots
  - ☐ Line Fit Plots
- Normal Probability:**
  - ☐ Normal Probability Plots

The 'OK' button is highlighted in blue.

Set data Level (from C5 to C 102) as Input Y Range, and set Index (from B5 to B 102) as Input X Range, then Click Residuals Button.

Click OK:

	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3	Regression Statistics								
4	Multiple R	0.521989							
5	R Square	0.272473							
6	Adjusted R Square	0.264894							
7	Standard Error	1.130287							
8	Observations	98							
9									
10	ANOVA								
11		df	SS	MS	F	Significance F			
12	Regression	1	45.93274	45.93274	35.95382	3.55E-08			
13	Residual	96	122.6446	1.277548					
14	Total	97	168.5774						
15									
16		Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
17	Intercept	10.20204	0.230111	44.33524	9.7E-66	9.745269	10.6588	9.745269	10.6588
18	X Variable	-0.0242	0.004036	-5.99615	3.55E-08	-0.03221	-0.01619	-0.03221	-0.01619
19									
20									
21									
22	RESIDUAL OUTPUT								
23									
24	Observation	Predicted Y	Residuals						
25	1	10.17784	0.202165						
26	2	10.15363	1.706366						
27	3	10.12943	0.840567						
28	4	10.10523	0.694768						
29	5	10.08103	-0.29103						
30	6	10.05683	0.33317						
31	7	10.03263	0.387371						
32	8	10.00843	0.811572						
33	9	9.984227	1.415773						
34	10	9.960026	1.359974						
35	11	9.935824	1.504176						
36	12	9.911623	1.768377						
37	13	9.887422	1.282578						
38	14	9.863221	0.666779						
39	15	9.83902	0.17098						
40	16	9.814819	0.095181						
41	17	9.790618	-0.65062						
42	18	9.766417	-0.60642						
43	19	9.742216	-0.19222						
44	20	9.718014	-0.04801						
45	21	9.693813	1.25381						
<div>Sheet1Lake HuronDeathsJeansRed Wine+</div>									

A new Sheet is created as show the result of regression analysis.

We can choose:

Regression ? X

Input

Input Y Range:

Input X Range:

☐ Labels ☐ Constant is Zero

☐ Confidence Level:  %

Output options

☒ Output Range:

☐ New Worksheet Ply:

☐ New Workbook

Residuals

☒ Residuals ☐ Residual Plots

☐ Standardized Residuals ☐ Line Fit Plots

Normal Probability

☐ Normal Probability Plots

OK Cancel Help

select a cell you want, and click ok:

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.521989							
R Square	0.272473							
Adjusted R Square	0.264894							
Standard Error	1.130287							
Observations	98							
ANOVA								
	df	SS	MS	F	Significance F			
Regression	1	45.93274	45.93274	35.95382	3.55E-08			
Residual	96	122.6446	1.277548					
Total	97	168.5774						
	Coefficient	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	10.20204	0.230111	44.33524	9.7E-66	9.745269	10.6588	9.745269	10.6588
X Variable 1	-0.0242	0.004036	-5.99615	3.55E-08	-0.03221	-0.01619	-0.03221	-0.01619
RESIDUAL OUTPUT								
Observation	Predicted Y	Residuals						
1	10.17784	0.202165						
2	10.15363	1.706366						
3	10.12943	0.840567						
4	10.10523	0.694768						
5	10.08103	-0.29103						
6	10.05683	0.33317						
7	10.03263	0.387371						
8	10.00843	0.811572						
9	9.984227	1.415773						
10	9.960026	1.359974						
11	9.935824	1.504176						
12	9.911623	1.768377						
13	9.887422	1.282578						
14	9.863221	0.666779						
15	9.83902	0.17098						
16	9.814819	0.095181						
17	9.790618	-0.65062						

The first part is regression statistics:

Regression Statistics	
Multiple R	0.521989
R Square	0.272473
Adjusted R Square	0.264894
Standard Error	1.130287
Observations	98

1. **Multiple R:** This is a value between -1 and 1, representing the degree of correlation between the independent variables and the dependent variable. Here, 0.52198923 indicates a moderate positive correlation between the independent variables and the dependent variable. The closer this value is to 1, the stronger the relationship
2. **R Square :** This is a measure of goodness-of-fit, indicating the proportion of variance in the dependent variable explained by the model. Here, 0.272472756 means the model can explain approximately 27.25% of the variance in the dependent variable. This implies that the model can account for some of the variability in the dependent variable, but there is still a substantial amount of unexplained variance.
3. **Adjusted R Square :** This is a modified version of R Square that takes into account the impact of the number of independent variables and the sample size on model performance. In this case, 0.264894347 is an adjusted R Square, providing a more accurate reflection of the model's fit.
4. **Standard Error :** This is the standard error of the model, indicating the average deviation between the actual values of the dependent variable and the model's predicted values. Here, 1.130286779 represents an average prediction error of approximately 1.13 units.
5. **Observations:** This represents the number of samples used to build the regression model, which is 98 in this example.

ANOVA					
	df	SS	MS	F	Significance F
Regression	1	45.93274	45.93274	35.95382	3.55E-08
Residual	96	122.6446	1.277548		
Total	97	168.5774			

The ANOVA (Analysis of Variance) summary table, commonly used in statistical analysis, particularly in the context of linear regression. Here's an explanation of each component of the table:

- **ANOVA:** This header indicates that the table is presenting an analysis of variance.
- **df (Degrees of Freedom):** This column shows the degrees of freedom associated with each component of the analysis. Degrees of freedom are a measure of the number of values in the final calculation of a statistic that are free to vary.
- **SS (Sum of Squares):** This column displays the sum of squared differences or variations. It represents the variability in the data that is attributed to each component.
- **MS (Mean Square):** This column represents the mean or average of the squared differences (SS) and is calculated by dividing SS by its respective degrees of freedom (df).

- **F (F-statistic):** The F-statistic is a test statistic used to compare the variances between groups or components. It is calculated by dividing the mean square of the regression (or model) by the mean square of the residuals (error term).
- **Significance F (Significance Level or p-value):** This column provides the p-value associated with the F-statistic. The p-value indicates the probability that the observed F-statistic occurred by chance. Smaller p-values typically suggest that the component being tested (in this case, the regression) is statistically significant.

Now, let's interpret the specific values in the table:

- **Regression:** This row represents the part of the variability in the dependent variable that is explained by the regression model. It has 1 degree of freedom, a sum of squares (SS) of 45.93274, a mean square (MS) of 45.93274, and an F-statistic of 35.95382. The very low p-value (3.55E-08) indicates that the regression model is highly significant in explaining the variation in the dependent variable.
- **Residual:** This row represents the unexplained or residual variability in the dependent variable. It has 96 degrees of freedom, an SS of 122.6446, and an MS of 1.277548.
- **Total:** This row represents the total variability in the dependent variable, combining both the explained (regression) and unexplained (residual) variation. It has 97 degrees of freedom and a total SS of 168.5774.

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	10.20204	0.230111	44.33524	9.7E-66	9.745269	10.6588	9.745269	10.6588
X Variable 1	-0.0242	0.004036	-5.99615	3.55E-08	-0.03221	-0.01619	-0.03221	-0.01619

This is a coefficients table from a regression analysis, providing information about the coefficients in the regression model and their statistical properties. Here's an explanation of each column in the table:

- **Coefficients:** This column displays the coefficients (parameter estimates) for each variable in the regression model, including the intercept and the independent variable X Variable 1. These coefficients represent the impact of each variable on the dependent variable.
- **Standard Error:** This column shows the standard error of the estimated coefficients, which measures the uncertainty in the coefficient estimates.
- **t Stat:** This column displays the t-statistic, which is a statistical measure used to test whether the coefficients are statistically significant. A higher absolute t-statistic (positive or negative) suggests greater statistical significance.
- **P-value:** This column provides the p-value associated with the t-statistic. The p-value indicates the probability of observing the t-statistic, and a smaller p-value suggests that the coefficient is statistically significant.
- **Lower 95% and Upper 95%:** These two columns show the lower and upper bounds of the 95% confidence interval for each coefficient. These intervals represent our level of confidence in the coefficient estimates.

Specifically:

- **Intercept:** The intercept is the constant term in the regression model, representing the predicted value of the dependent variable when the independent variable is 0. In this case, the estimated intercept is 10.20203661, with a very small standard error (0.230111251). The



t-statistic is very large (44.33523595), and the corresponding p-value (9.70257E-66) is very close to zero, indicating that the intercept is highly significant.

- **X Variable 1:** The coefficient estimate for X Variable 1 is -0.024201111, with a standard error of 0.004036108. The t-statistic is -5.99615055, and the associated p-value (3.54523E-08) is very close to zero, indicating that X Variable 1 has a statistically significant effect on the dependent variable.

The last part is residual output:

RESIDUAL OUTPUT		
Observation	Predicted Y	Residuals
1	10.17784	0.202165
2	10.15363	1.706366
3	10.12943	0.840567
4	10.10523	0.694768
5	10.08103	-0.29103
6	10.05683	0.33317
7	10.03263	0.387371
8	10.00843	0.811572
9	9.984227	1.415773
10	9.960026	1.359974
11	9.935824	1.504176
12	9.911623	1.768377
13	9.887422	1.282578
14	9.863221	0.666779
15	9.83902	0.17098
16	9.814819	0.095181
17	9.790618	-0.65062
18	9.766417	-0.60642

It shows the predicted value of the regression model and the residuals of predicted values and original values.

"Residual" is an important concept in statistics and regression analysis, which represents the difference or error between the actual observed value and the predicted value of the model.

In regression analysis, we try to build a mathematical model to explain the relationship between the dependent variable (or response variable) and one or more independent variables (or explanatory variables). The difference between the model's predicted value and the actual observed value is the residual. Residuals represent the variability in the dependent variable that is not explained or captured by the model. Our goal is to minimize the residual error so that it is as close to zero as possible, thus improving the fit of the model.

Residuals are usually calculated in the following way:

1. First, use a regression model to calculate the predicted value for each observation.
2. Then, subtract the corresponding predicted value from the actual value of each observation to obtain the residual.

The nature of the residuals is very important in assessing the quality and accuracy of the regression model. If the residuals have a certain pattern or structure, it may mean there is a problem with the model and further improvement is needed. Some common residual patterns include:

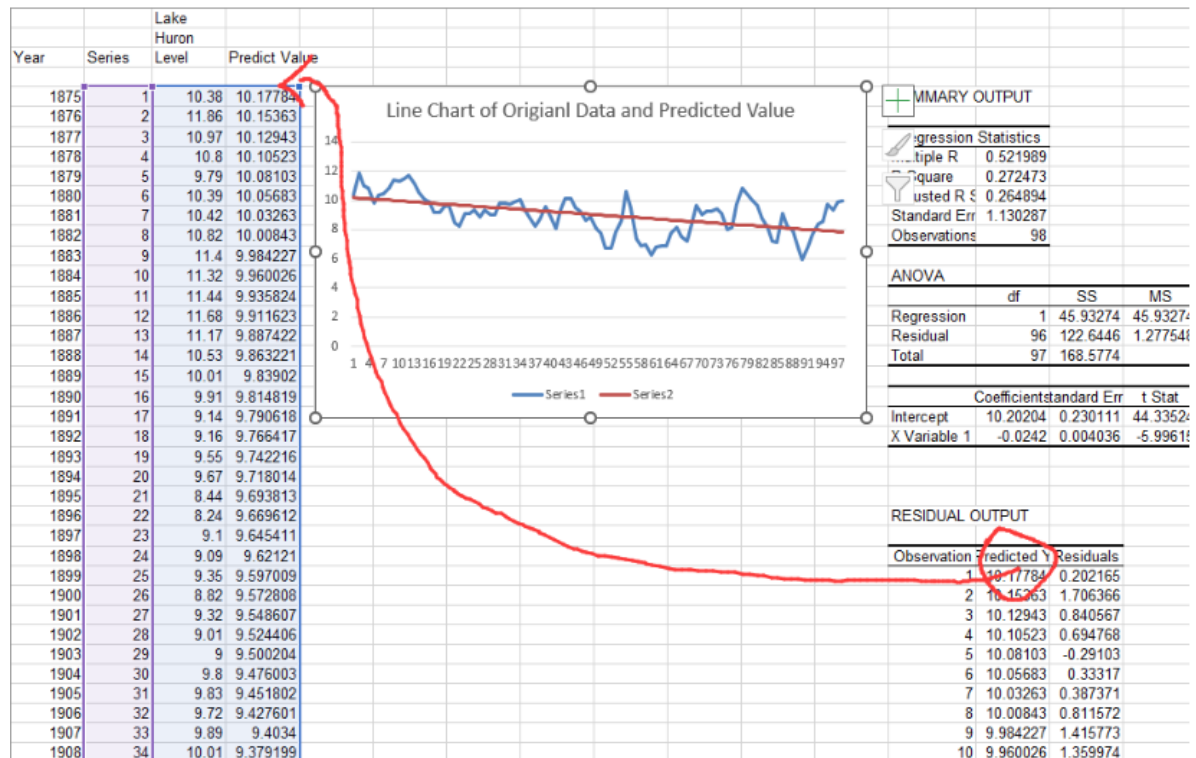
- **Random Residuals:** Residuals are randomly distributed with no obvious pattern or trend.
- **Heteroscedastic Residuals:** The variance of the residuals is inconsistent at different levels of the independent variables, and transformation or weighted regression may need to be

considered.

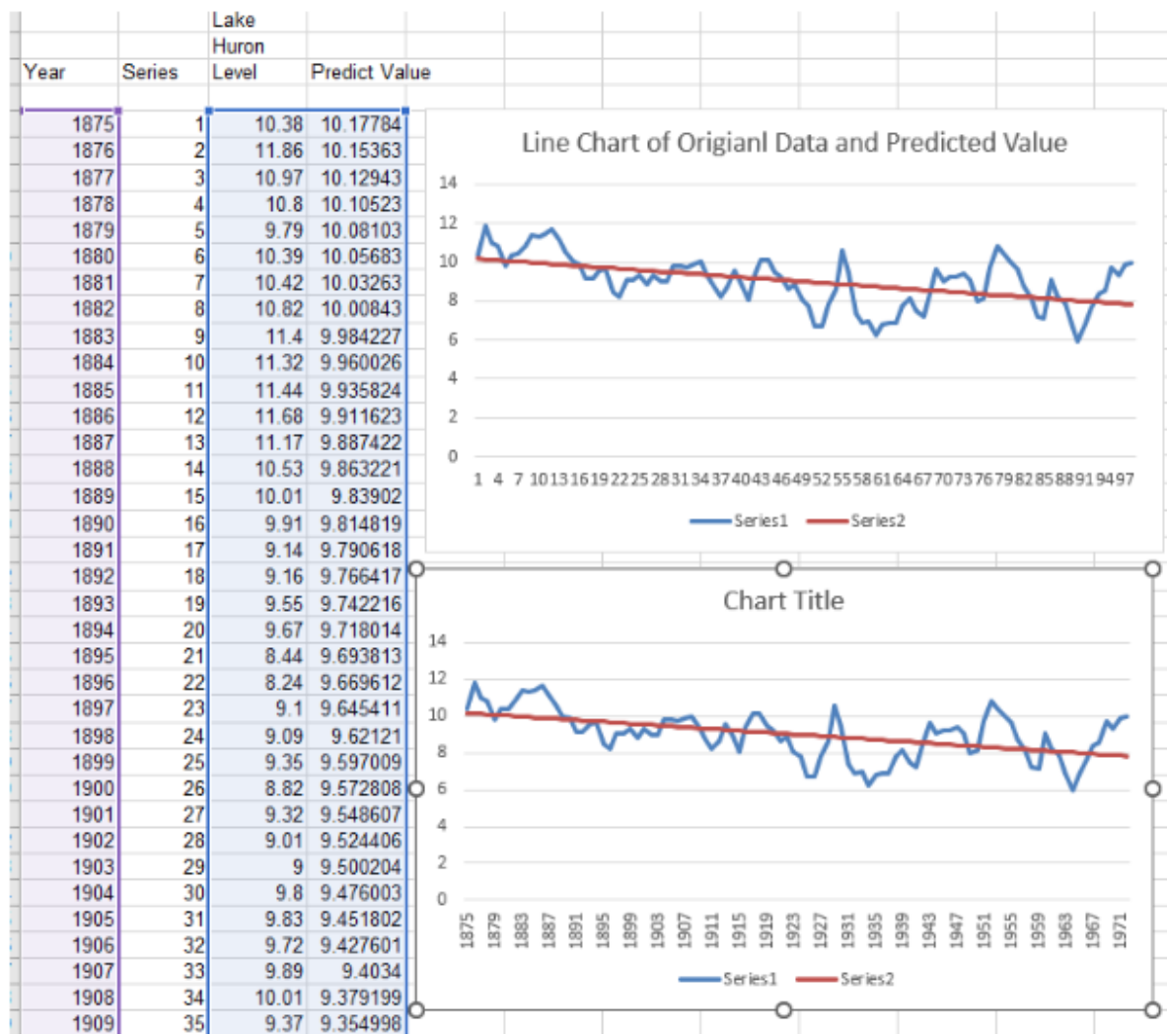
- **Autocorrelated residuals:** There is temporal or spatial correlation between the residuals, which needs to be processed using a time series or spatial model.
- **Nonlinear Residuals:** There is a nonlinear relationship between the residuals and the independent variables, which may require a more complex model.

By analyzing the residuals, model deficiencies can be identified and appropriate actions can be taken to improve the model's fit and predictive performance. Therefore, residuals are a key concept in regression analysis and help in evaluating the effectiveness of a model.

Now copy the predicted values from result output to the beside original data:



Or set the year as x value but just index:



Judging from the visualization results of the predicted values of the regression model, there is a downward trend. But it can be argued that the model does not fit the trend of the original data very well.

Try quadratic to fit the data:

C3						
fx						
Series * 2						
	A	B	C	D	E	F
1				Lake		
2				Huron		
3	Year	Series	Series * 2	Level	Predict Value	
4						
5	1875	1	1	10.38	10.17784	
6	1876	2	4	11.86	10.15363	
7	1877	3	9	10.97	10.12943	
8	1878	4	16	10.8	10.10523	
9	1879	5	25	9.79	10.08103	

Now try regression again:

	Series	Series * 2	Lake Huron Level	Predict Va	Predicted Values of X^2
875	1	1	10.38	10.17784	11.22617
876	2	4	11.86	10.15363	11.13712
877	3	9	10.97	10.12943	11.04943
878	4	16	10.8	10.10523	10.96308
879	5	25	9.79	10.08103	10.87809
880	6	36	10.39	10.05683	10.79444
881	7	49	10.42	10.03263	10.71215
882	8	64	10.82	10.00843	10.63121
883	9	81	11.4	9.984227	10.55162
884	10	100	11.32	9.960026	10.47338
885	11	121	11.44	9.935824	10.3965
886	12	144	11.68	9.911623	10.32096
887	13	169	11.77	9.887422	10.24677
888	14	196	10.53	9.863221	10.17394
889	15	225	10.01	9.83902	10.10245
890	16	256	9.91	9.814819	10.03232
891	17	289	9.14	9.790618	9.963538
892	18	324	9.16	9.766417	9.896107
893	19	361	9.55	9.742216	9.830027
894	20	400	9.67	9.718014	9.765297
895	21	441	8.44	9.693813	9.701919
896	22	484	8.24	9.669612	9.639891
897	23	529	9.1	9.645411	9.579215

**Regression**

Input

Input Y Range:

Input X Range:

☐ Labels ☐ Constant is Zero

☐ Confidence Level:  %

Output options:

☒ Output Range:

☐ New Worksheet Ply:

☐ New Workbook

Residuals

☒ Residuals ☐ Residual Plots

☐ Standardized Residuals ☐ Line Fit Plots

Normal Probability

☐ Normal Probability Plots

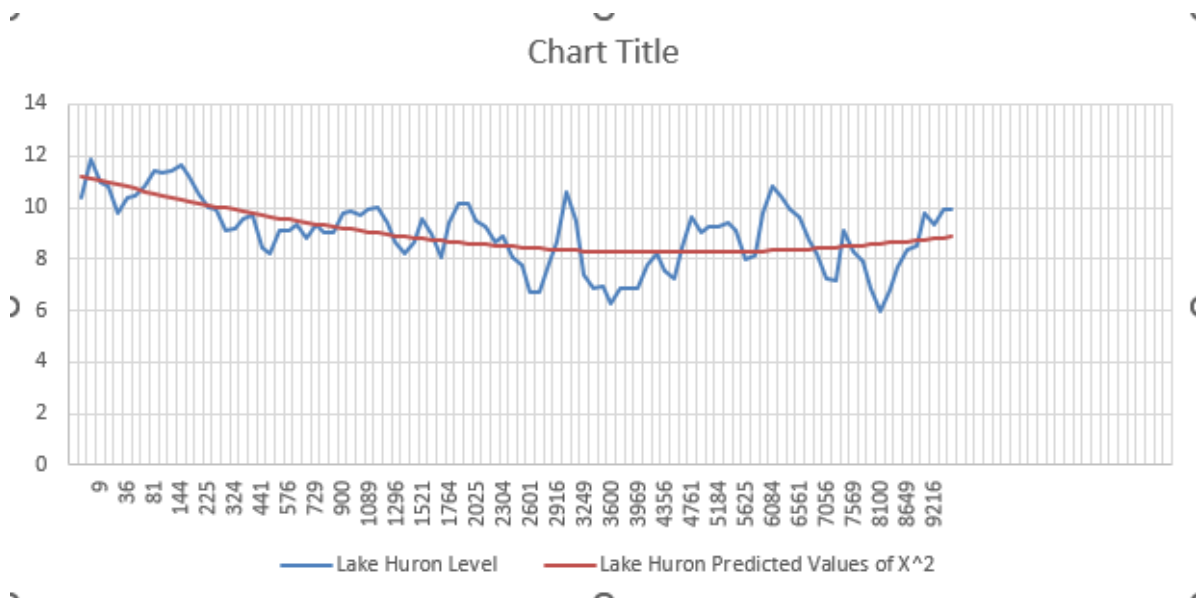
Now we get the regression result:

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.638998							
R Square	0.408319							
Adjusted R Square	0.395862							
Standard Error	1.024665							
Observations	98							
ANOVA								
	df	SS	MS	F	Significance F			
Regression	2	68.83327	34.41664	32.77969	1.49E-11			
Residual	95	99.7441	1.049938					
Total	97	168.5774						
	Coefficient	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	11.31656	0.316967	35.70265	7.58E-57	10.68731	11.94582	10.68731	11.94582
X Variable 1	-0.09107	0.014779	-6.16242	1.72E-08	-0.12041	-0.06173	-0.12041	-0.06173
X Variable 2	0.000675	0.000145	4.670259	9.88E-06	0.000388	0.000963	0.000388	0.000963
RESIDUAL OUTPUT								
Observation	Predicted Y	Residuals						
1	11.22617	-0.84617						
2	11.13712	0.722879						
3	11.04943	-0.07943						
4	10.96308	-0.16308						
5	10.87809	-1.08809						
6	10.79444	-0.40444						
7	10.71215	-0.29215						
8	10.63121	0.188788						
9	10.55162	0.848377						
10	10.47338	0.846616						
11	10.3965	1.043504						
12	10.32096	1.359041						
13	10.24677	0.923227						
14	10.17394	0.356062						
15	10.10245	-0.09245						
16	10.03232	-0.12232						
17	9.963538	-0.82354						
18	9.896107	-0.73611						
19	9.830027	-0.280027						

Then we can copy predicted to the beside of original data again:

	Lake Huron			
Series * 2	Level	Predict Va	Predicted Values of X^2	
1	10.38	10.17784	11.22617	
4	11.86	10.15363	11.13712	
9	10.97	10.12943	11.04943	
16	10.8	10.10523	10.96308	
25	9.79	10.08103	10.87809	
36	10.39	10.05683	10.79444	
49	10.42	10.03263	10.71215	
64	10.82	10.00843	10.63121	
81	11.4	9.984227	10.55162	
100	11.32	9.960026	10.47338	
121	11.44	9.935824	10.3965	
144	11.68	9.911623	10.32096	
169	11.17	9.887422	10.24677	
196	10.53	9.863221	10.17394	
225	10.01	9.83902	10.10245	
256	9.91	9.814819	10.03232	

Then we can graph that:



This seems much better, and judging from the trend of the model, the model seems to show an upward trend in the later period.

Check R-square values:

SUMMARY OUTPUT											SUMMARY OUTPUT	
Regression Statistics											Regression Statistics	
Multiple R	0.521989										Multiple R	0.638998
R Square	0.272473										R Square	0.408319
Adjusted R Square	0.264894										Adjusted R Square	0.395862
Standard Error	1.130287										Standard Error	1.024665
Observations	98										Observations	98

The two R Square values are:

1. The R Square value of the first model is 0.272472756.
2. The R Square value for the second model is 0.408318578.

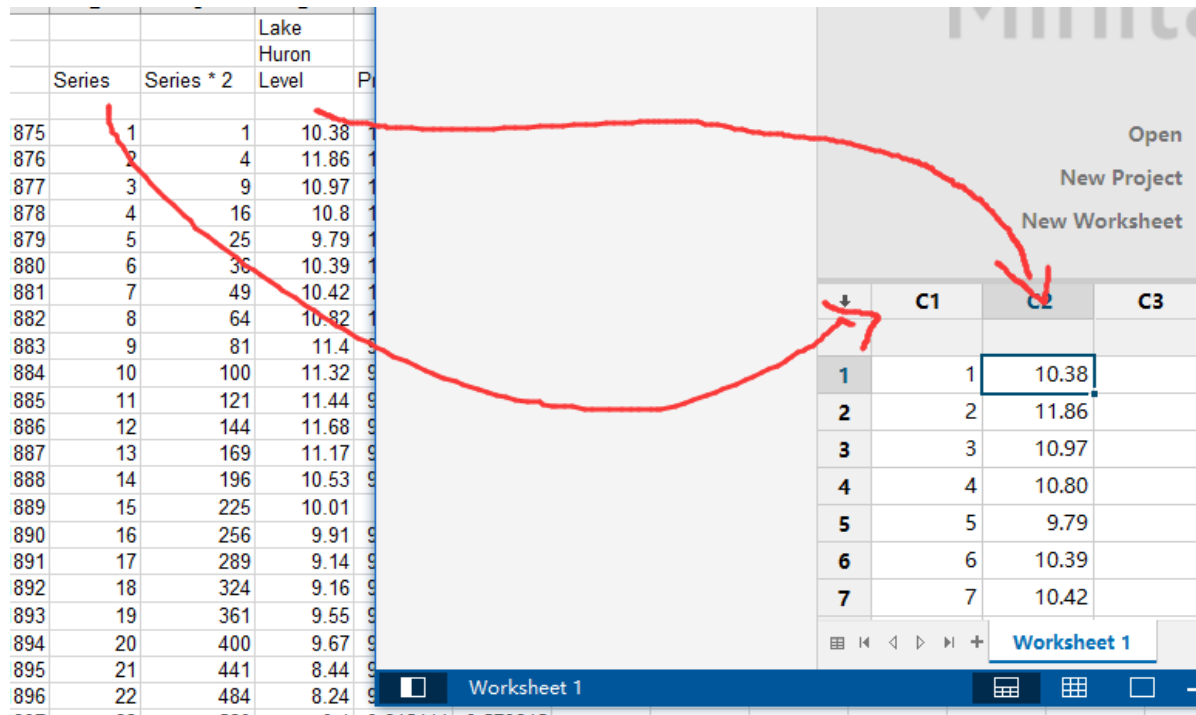
These two R Square values measure how well the regression model explains the dependent variable. When comparing these two values, you can see:

- The second model has a higher R Square value of 0.408318578, which is closer to 1 than the first model's 0.272472756. This means that the second model better explains the variability in the dependent variable and has a better fit.
- Since R Square reflects the goodness of fit of the model, higher R Square values generally indicate a better fit of the model to the data. Therefore, the second model is more powerful in explaining the data.

Overall, the R Square value of the second model indicates that it has better fitting performance relative to the first model and is better able to explain the variability of the dependent variable.

## Minitab

Copy X and Y to the minitab:

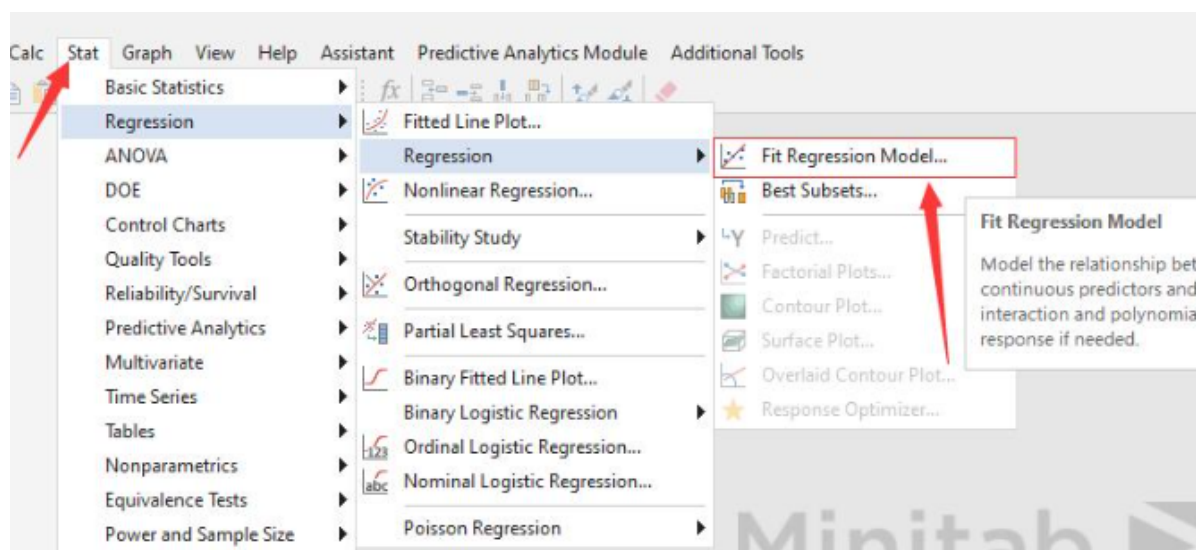


	Series	Series * 2	Lake Huron Level	P
875	1	1	10.38	1
876	2	4	11.86	1
877	3	9	10.97	1
878	4	16	10.8	1
879	5	25	9.79	1
880	6	36	10.39	1
881	7	49	10.42	1
882	8	64	10.82	1
883	9	81	11.4	9
884	10	100	11.32	9
885	11	121	11.44	9
886	12	144	11.68	9
887	13	169	11.17	9
888	14	196	10.53	9
889	15	225	10.01	9
890	16	256	9.91	9
891	17	289	9.14	9
892	18	324	9.16	9
893	19	361	9.55	9
894	20	400	9.67	9
895	21	441	8.44	9
896	22	484	8.24	9

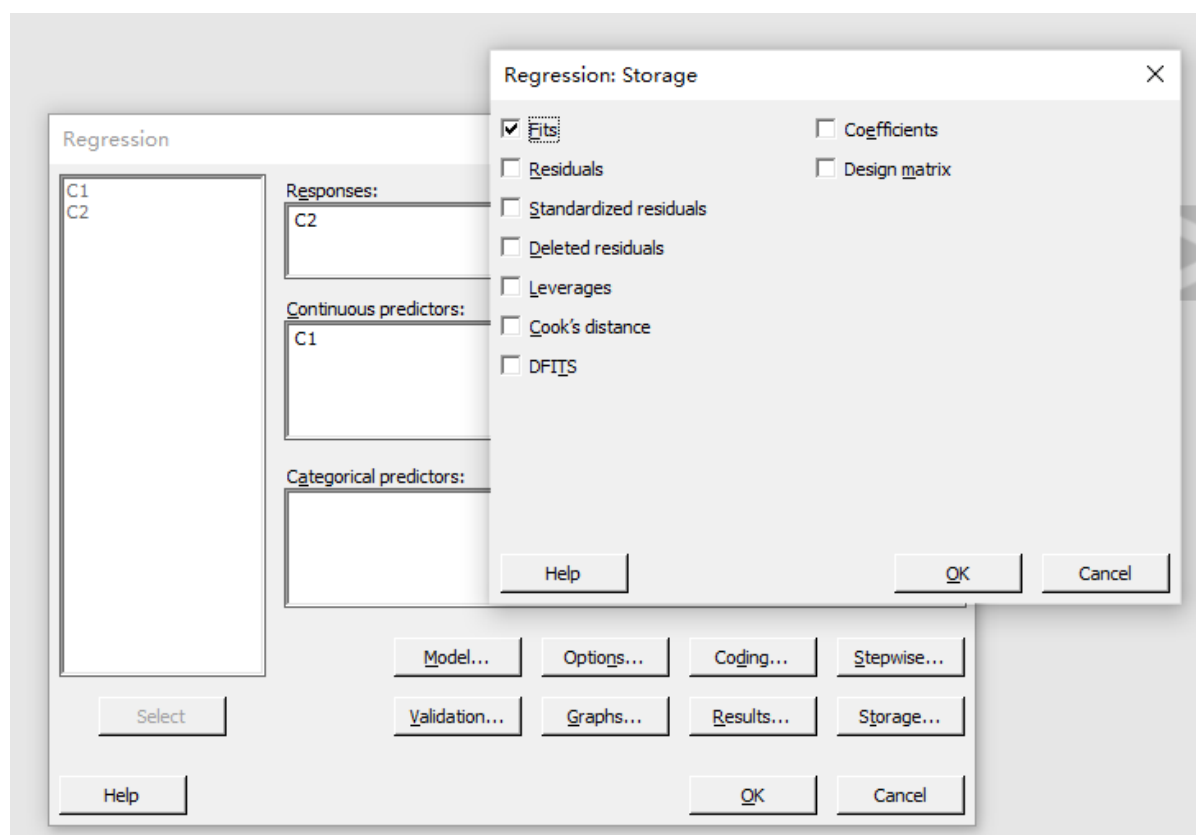
	C1	C2	C3
1	1	10.38	
2	2	11.86	
3	3	10.97	
4	4	10.80	
5	5	9.79	
6	6	10.39	
7	7	10.42	

Then Click Statistics → Regression → Regression:

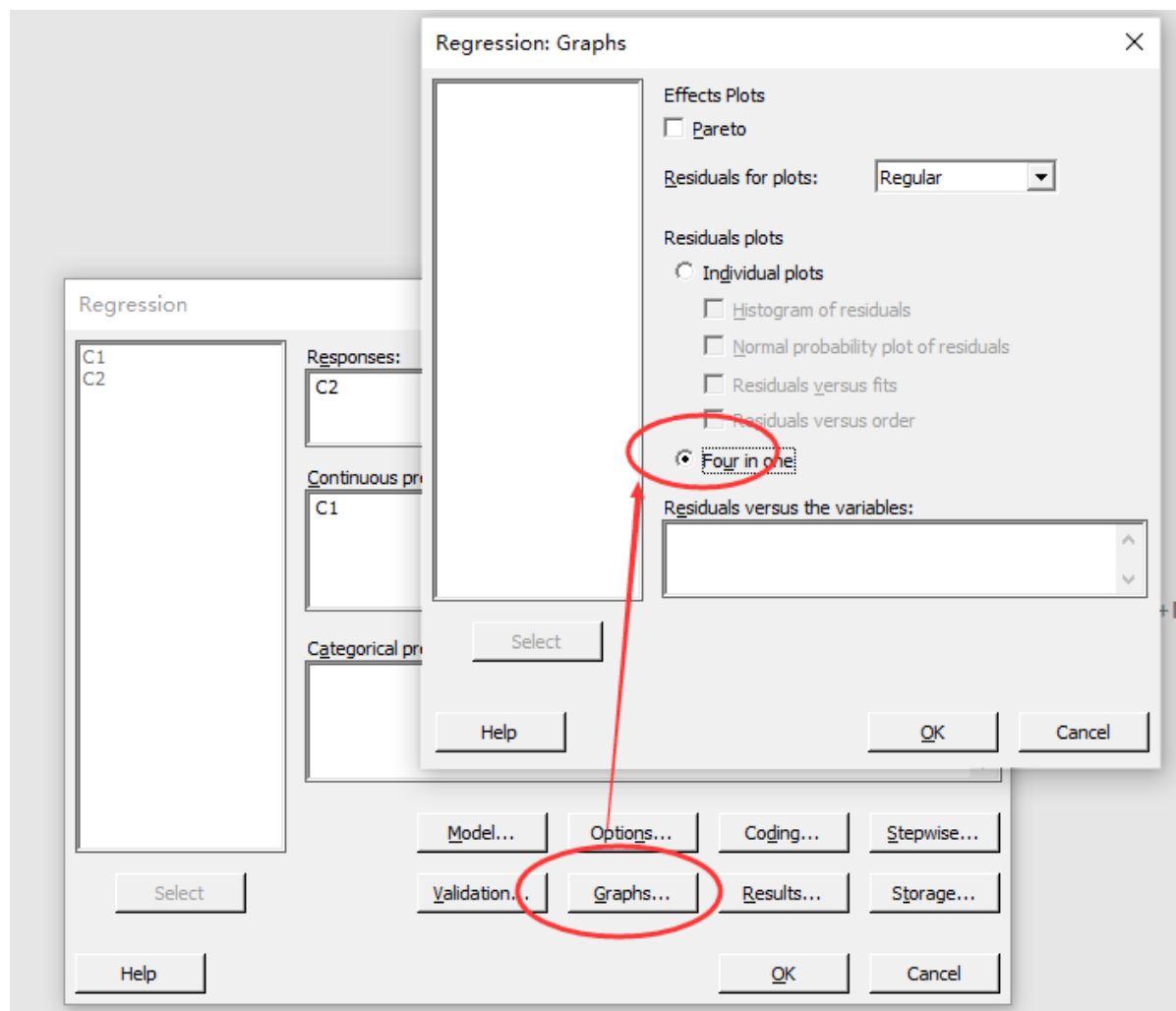


Select Columns:

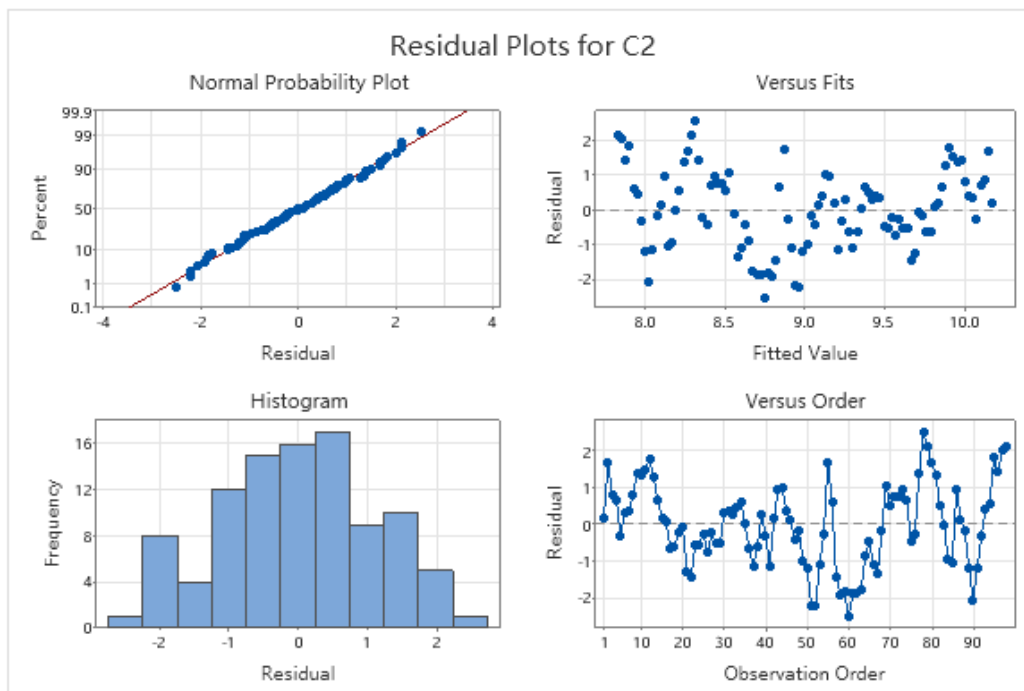




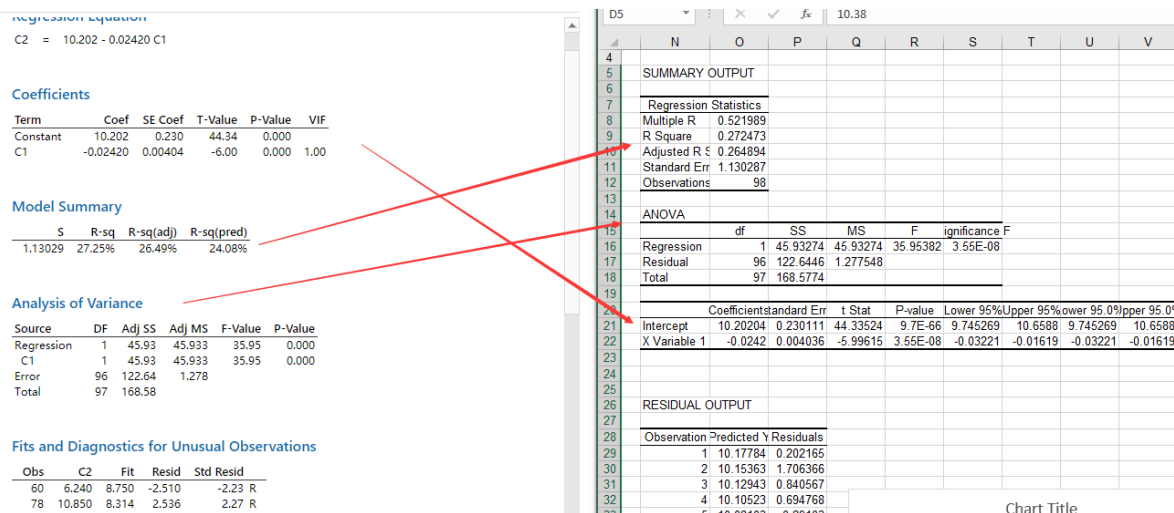
The click graphs and click four in one:



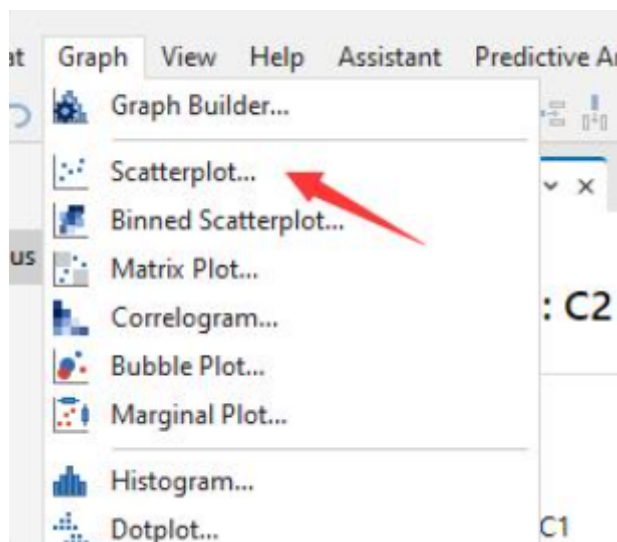
Then a graph is created:



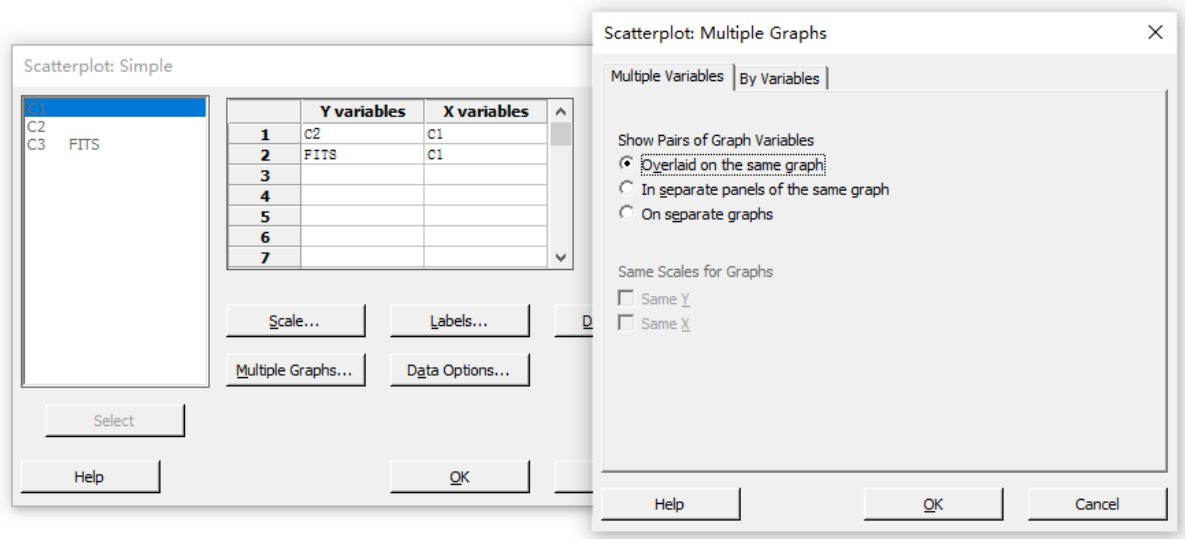
We can compare excel and minitab results:



The go to graph and select scatter plot:

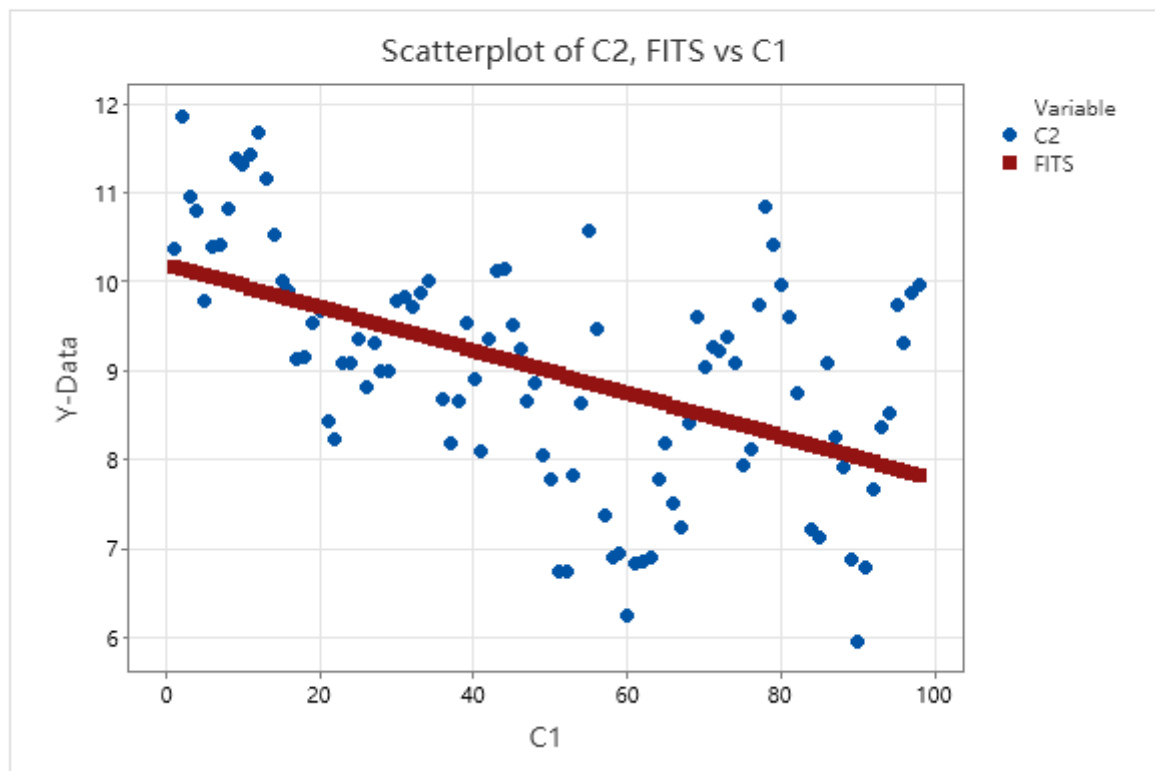






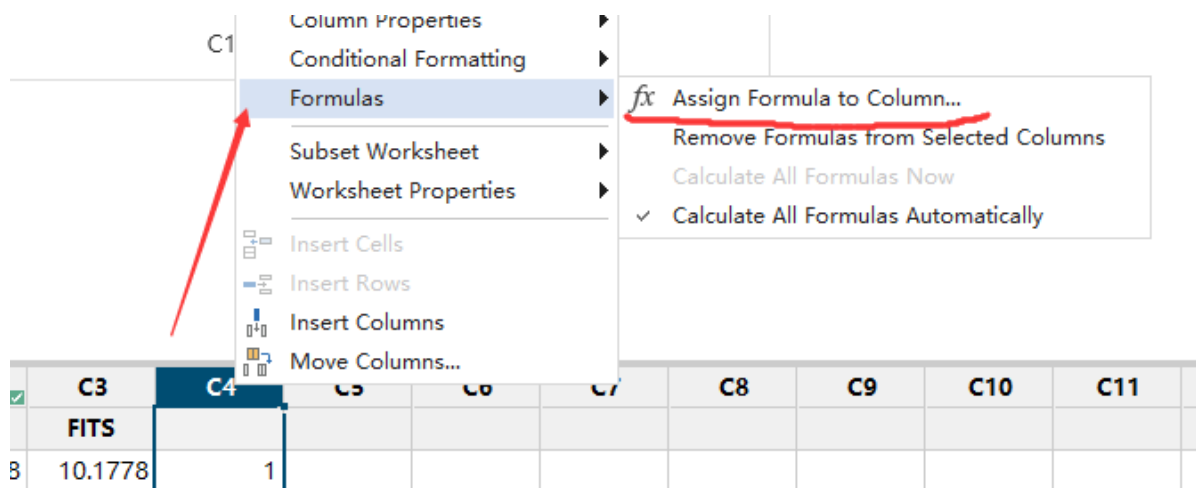
The that is the result:

## Scatterplot of C2, FITS vs C1

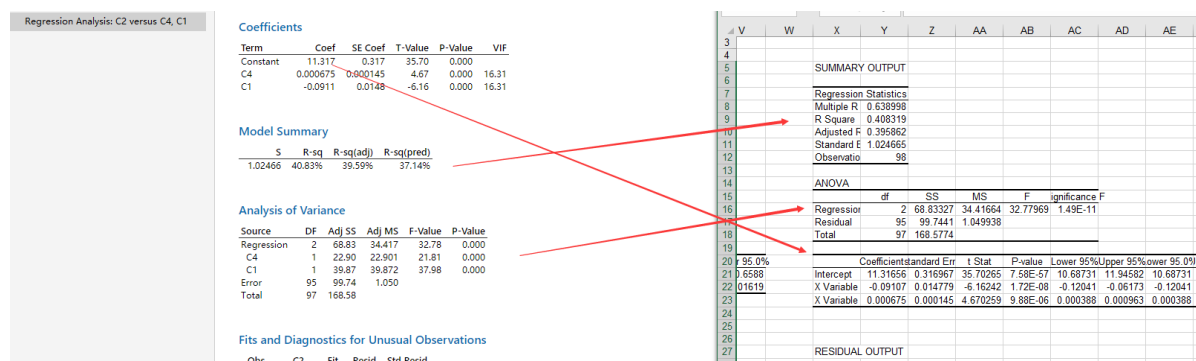
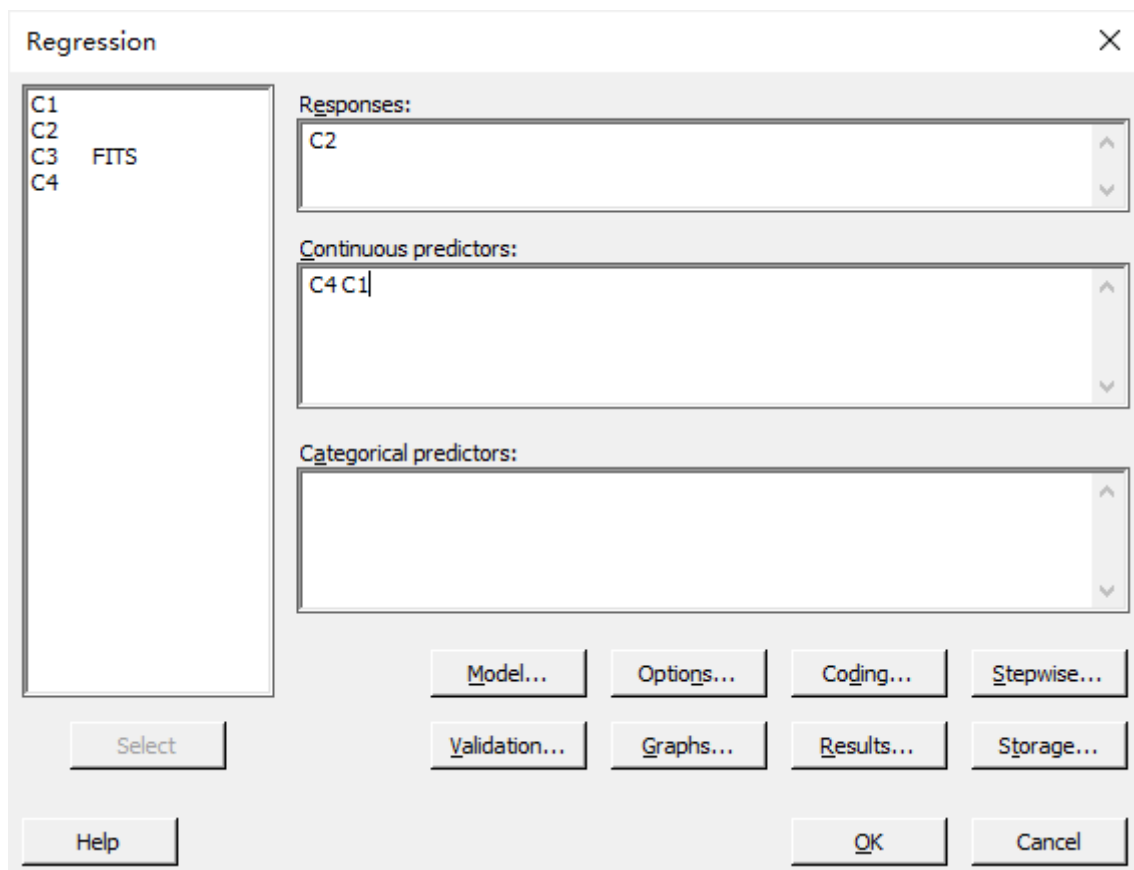


Obviously this is not the best model, no matter from the visualization results or from the value of  $r$  square.

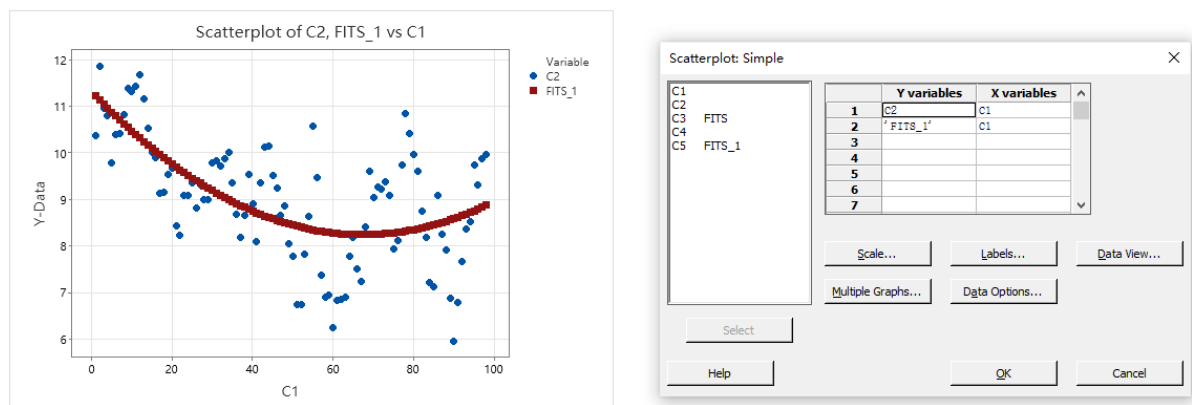
We can try  $x^2$  next:



Then:



Now we use scatter plot graph to draw that again:



## Additional Information

The significance of using the square of  $x$  ( $X^2$ ) and  $x$  as the  $x$ -axis is to introduce a quadratic term, thus extending linear regression into quadratic regression or polynomial regression. This can be used to model non-linear relationships, allowing the model to fit the data better, especially when the data shows a clear curvilinear trend.

Specifically, if you include  $x$  and  $X^2$  as the two features of the  $x$ -axis, your regression model becomes:

$$y = \beta_0 + \beta_1 * x + \beta_2 * X^2 + \varepsilon$$

in that formula:

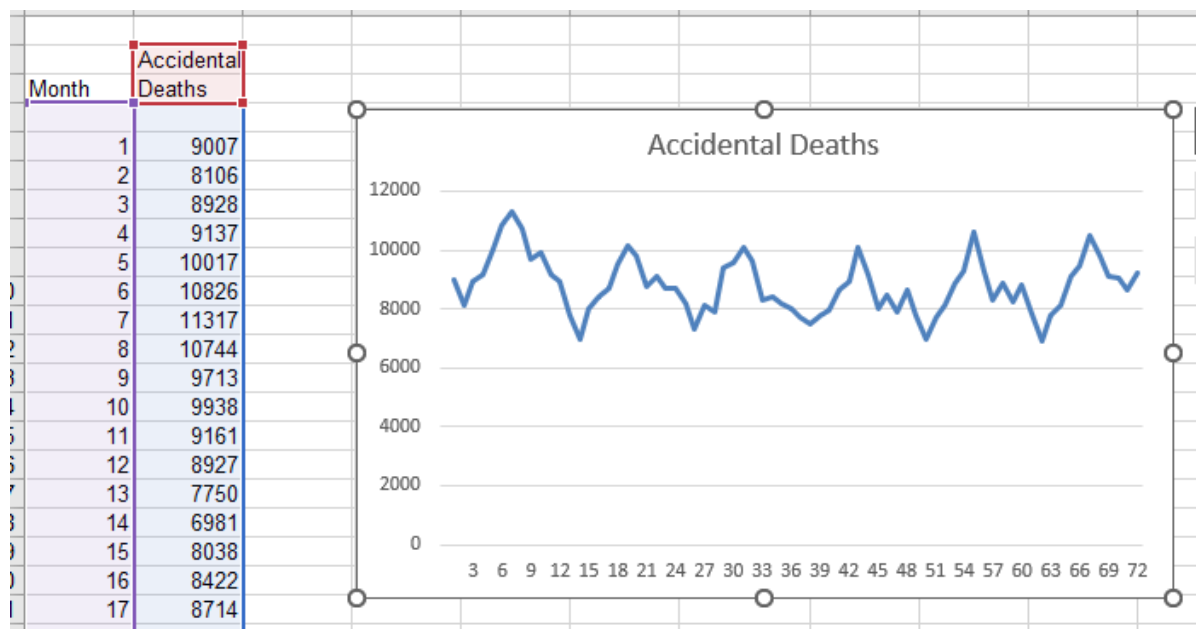
- $y$  is the dependent variable (usually the value you want to predict).
- $x$  is the original independent variable (year).
- $X^2$  is the square of  $x$ .
- $\beta_0$ ,  $\beta_1$ , and  $\beta_2$  are the regression coefficients used to fit the model.
- $\varepsilon$  is the error term.

The point of this model is that it allows you to capture the non-linear relationship between the dependent variables  $y$  and  $x$ . If your data shows a curvilinear trend, a quadratic regression model may fit the data better and provides more flexibility than a simple linear regression model.

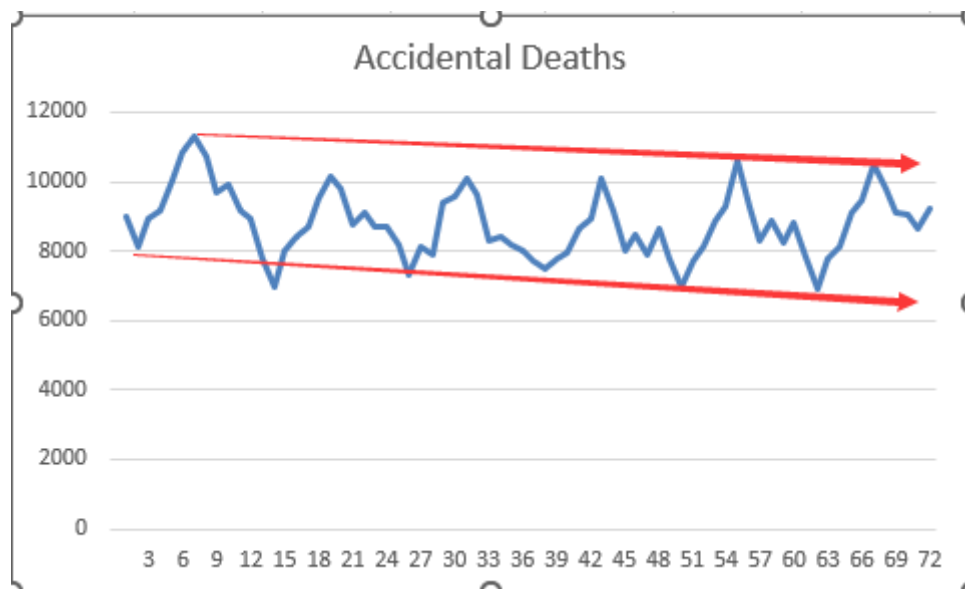
However, it should be noted that introducing quadratic terms may also increase the risk of overfitting, so the complexity of the model needs to be carefully selected and whether it is necessary to introduce quadratic terms or higher-order terms based on the characteristics of the data. Often, you can use model evaluation metrics (such as mean squared error, R squared, etc.) to help determine the most suitable model.

## Death

For death, it is same to draw the graph first.

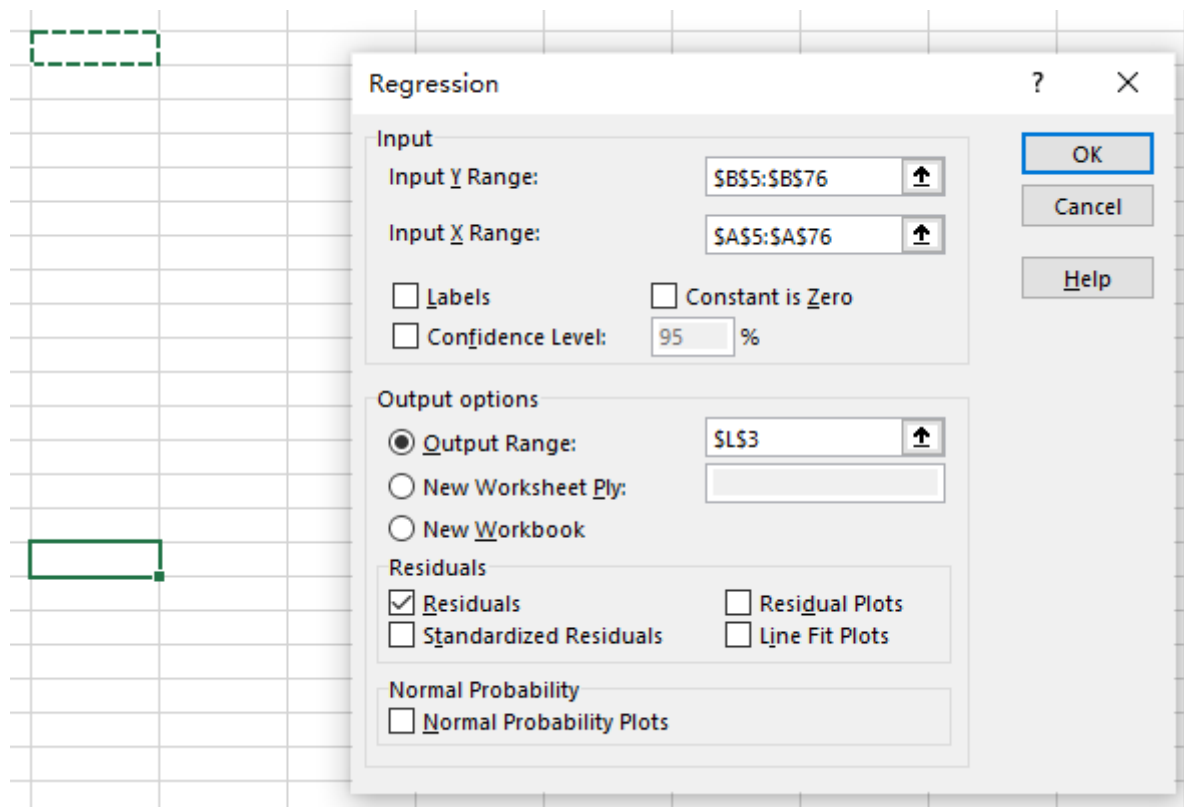


If we observe directly, we can think that the trend is showing a slowly decreasing trend:



In addition, the data should have seasonality.

and then we can do the regression again:



That is the result:

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.18517							
R Square	0.034288							
Adjusted R Square	0.020492							
Standard Error	948.468							
Observations	72							
ANOVA								
	df	SS	MS	F	Significance F			
Regression	1	2235820	2235820	2.485372	0.119418			
Residual	70	62971414	899591.6					
Total	71	65207234						
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	9097.225	225.9052	40.27009	3.75E-50	8646.671	9547.779	8646.671	9547.779
X Variable	-8.47915	5.378442	-1.57651	0.119418	-19.2061	2.247817	-19.2061	2.247817

The above output is a summary table of the results of a regression analysis, often used to explain the performance and parameter estimates of regression models. Below I'll explain what each part means:

### Regression Statistics:

1. **Multiple R:** Multiple correlation, indicating the linear correlation between the dependent variable (dependent variable) and the independent variable (independent variable). Here, Multiple R is equal to 0.185, indicating a weak linear correlation.
2. **R Square:** Coefficient of determination, indicating how well the model fits. It is the proportion of the variation in the dependent variable that can be explained by the independent variables. Here, R Square is 0.034, indicating that the model is able to explain approximately 3.4% of the variation in the dependent variable.

3. **Adjusted R Square:** The adjusted coefficient of determination, which takes into account the number of independent variables and sample size, is used to more accurately estimate the fit of the model. Here, the corrected R Square is 0.0205.
4. **Standard Error:** Standard error, used to measure the error of the regression item in the model. Here, the standard error is 948.47.
5. **Observations:** The number of observations, indicating the number of data points used to fit the regression model, here are 72 observations.

#### ANOVA table (Analysis of Variance):

This table shows the variance decomposition of the regression analysis, which helps us understand the statistical significance of the model.

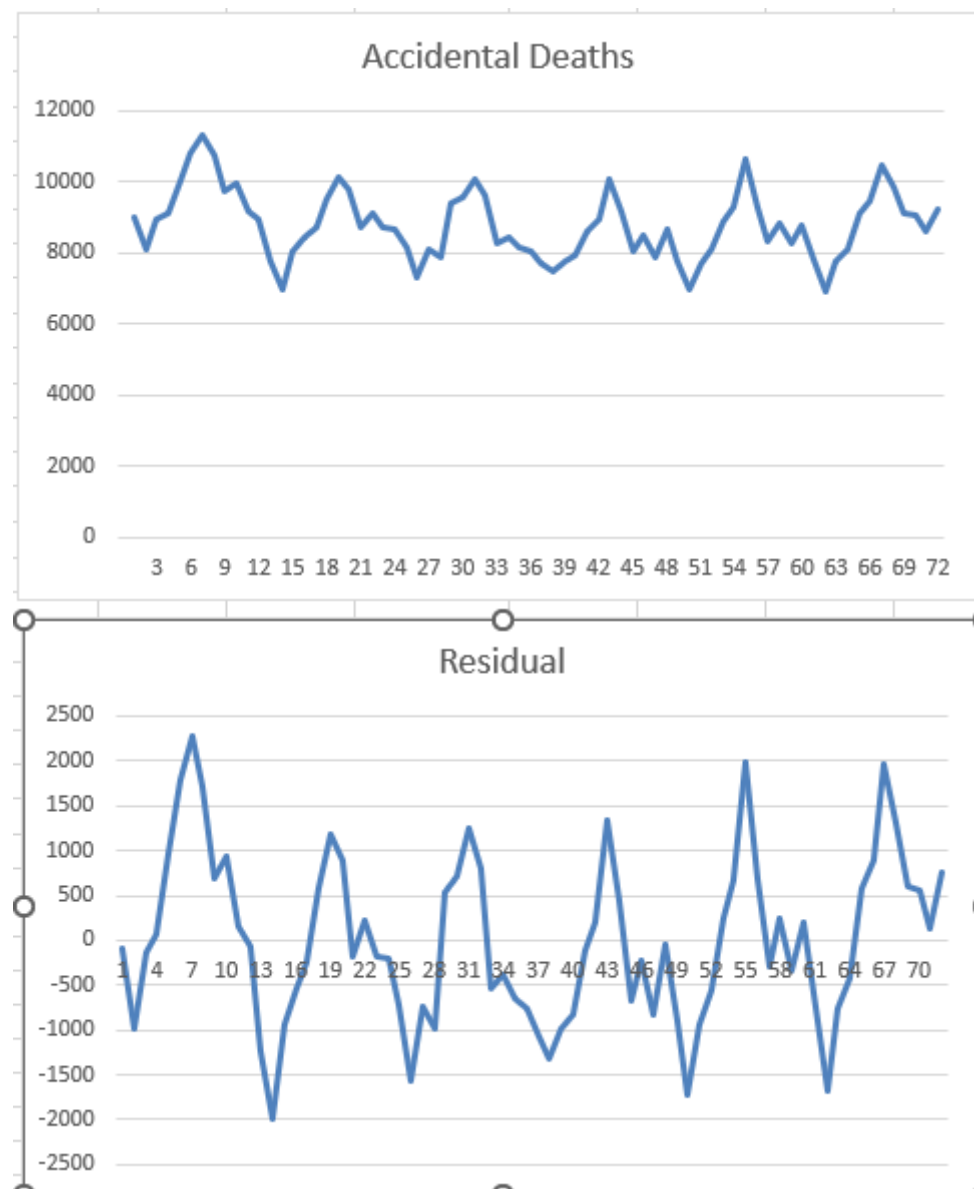
1. **df:** degrees of freedom, indicating the number of parameters that can be changed freely in the model. Here, the regression degree of freedom is 1 and the residual degrees of freedom are 70.
2. **SS:** sum of squares, indicating the sum of squares corresponding to each variable. Here, the regression sum of squares is 2235819.523 and the residual sum of squares is 62971414.46.
3. **MS:** Mean Square, which is the sum of squares divided by the degrees of freedom, used to calculate the F-statistic.
4. **F:** F statistic, used to test the significance of the regression model. Here, the F statistic is 2.49.
5. **Significance F:** Significance level, indicating the significance of the F statistic. Here, the p-value (Significance F) is 0.1194, which is greater than the commonly used significance level (for example, 0.05), indicating that the model is not statistically significant.

#### Coefficient table:

This table lists the regression model's coefficients (Intercept and X Variable 1) and statistics associated with them.

1. **Intercept:** Intercept, indicating the estimated value of the dependent variable when the independent variable X Variable 1 is equal to zero. Here, the intercept is 9097.22.
2. **X Variable 1:** independent variable 1, its coefficient is -8.48. This means that, holding other variables constant, each unit increase in X Variable by 1 causes a change in the dependent variable.
3. **Standard Error:** Standard error, used to estimate the uncertainty of the coefficient.
4. **t Stat:** t statistic, used to test whether the coefficient is significantly different from zero. Here, the t-statistic for X Variable 1 is -1.58, which corresponds to a p-value of 0.1194, so X Variable 1 is not statistically significant.
5. **P-value:** p value, used to test whether the coefficient is significantly different from zero. Generally, a coefficient is considered significant if the p-value is less than a selected significance level (such as 0.05).
6. **Lower 95% and Upper 95%:** These columns show the confidence interval for each coefficient, indicating the estimated range of the coefficient.

Based on this information, you can conclude that the model has a low R Square, the F statistic is not significant, and the coefficient of X Variable 1 is not significant. This may mean that the model does not explain the variation in the dependent variable well, or that the model needs to be further improved to better fit the data.



Visualize the residuals, which also seem to have seasonality.

However, due to the poor performance of the model, we will try multiple regression first.

	A	B	C
			Accidental
	Month	Month^2	Deaths
	1	1	9007
	2	4	8106
	3	9	8928
	4	16	9137
	5	25	10017
	6	36	10826
	7	49	11317
	8	64	10744
	9	81	9713
	10	100	9938

Regression

Input

Input Y Range:

SC\$5:SC\$76

↑

Input X Range:

SA\$5:SB\$76

↑

☐ Labels

☐ Constant is Zero

☐ Confidence Level:

95

%

Output options

☒ Output Range:

\$W\$3

↑

☐ New Worksheet Ply:

☐ New Workbook

Residuals

☒ Residuals

☐ Residual Plots

☐ Standardized Residuals

☐ Line Fit Plots

Normal Probability

☐ Normal Probability Plots

OK

Cancel

Help

SUMMARY OUTPUT

Regression Statistics									
Multiple R	0.369468								
R Square	0.136507								
Adjusted R Square	0.111478								
Standard Error	903.3432								
Observations	72								
ANOVA									
	df	SS	MS	F	Significance F				
Regression	2	8901238	4450619	5.453997	0.006323				
Residual	69	56305996	816028.9						
Total	71	65207234							
		Coefficient	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	9806.528	328.4618	29.85592	3.49E-41	9151.264	10461.79	9151.264	10461.79	
X Variable 1	-65.9902	20.76465	-3.178	0.00222	-107.415	-24.5658	-107.415	-24.5658	
X Variable 2	0.787822	0.275656	2.857992	0.005632	0.237904	1.337741	0.237904	1.337741	

This new regression model performs better relative to the previous model. Here are the key observations about the new model:

### Regression Statistics:

1. **Multiple R:** The multiple correlation has improved from about 0.185 before to about 0.369, which means a stronger linear correlation between the independent variable and the dependent variable.
2. **R Square:** The coefficient of determination has improved from about 0.034 before to about 0.137, which means that the new model better explains about 13.7% of the variation in the dependent variable.
3. **Adjusted R Square:** The adjusted coefficient of determination has improved from about 0.0205 before to about 0.1115, which shows that the new model also performs better when considering the number of independent variables and sample size.
4. **Standard Error:** The standard error has been reduced from the previous about 948.47 to about 903.34, which means that the fitting error of the model is smaller and the model is more accurate.
5. **ANOVA table:** The F statistic here is 5.45, and the significance level (Significance F) is 0.0063, which shows that the regression model is statistically significant, which is a significant difference compared with the previous model Improve.

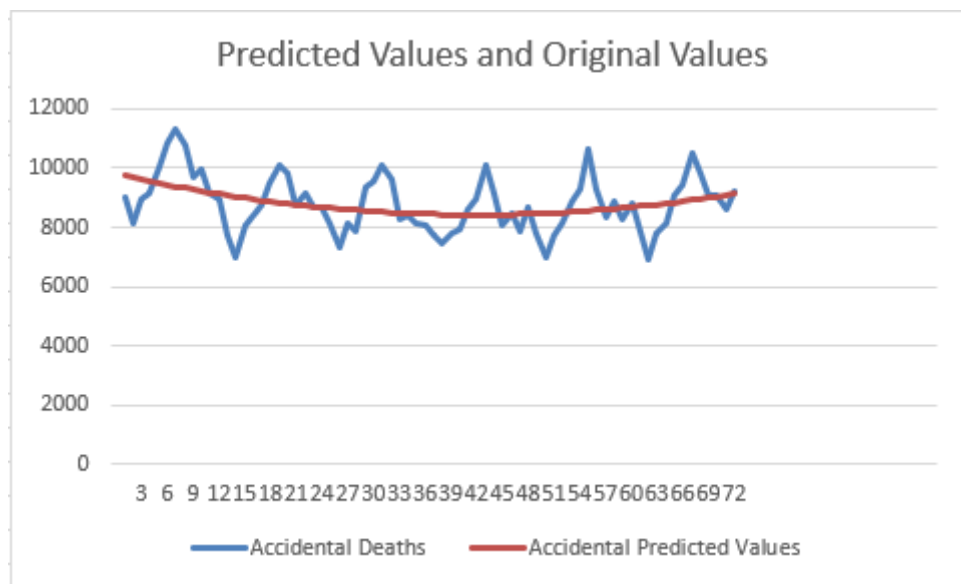
### Coefficient table:

1. **Intercept:** The intercept is still significant, its t-statistic is much greater than zero, and the p-value is very small.
2. **X Variable 1:** The coefficient of X Variable 1 becomes -65.99, which means that each unit increase of X Variable 1 will lead to a decrease of the dependent variable, which is more effective than the previous model The impact is clear and significant.
3. **X Variable 2:** The new model introduces an additional independent variable, X Variable 2, which has a coefficient of 0.79, a t-statistic of 2.86, and a p-value of 0.0056, indicating that X Variable 2 is also significant.

To summarize, the new regression model performs better relative to the previous model. It has a higher coefficient of determination, lower standard errors, and more significant F-statistics and coefficients. This means that the new model better explains the changes in the dependent variable and fits the data more accurately. Therefore, the new model is an improved model from a statistical and interpretive performance perspective.

Then created a visualization to show the predicted values and original values:





Compare two residuals:

