Chapter 4 Data Browser

Data Browser

Data Browser includes data exploration and basic analysis

Node using Data Browser

Data Browser can be used in the following Input/Output Nodes:

No.	Category	Name
1	Input Node	File Reader, File Reader2, Excel Reader, Copy&Paste Input, Access Reader, ODBC Reader, OLEDB Reader, Oracle Reader
2	Output Node	Display Node, Result Pivoting

Getting started

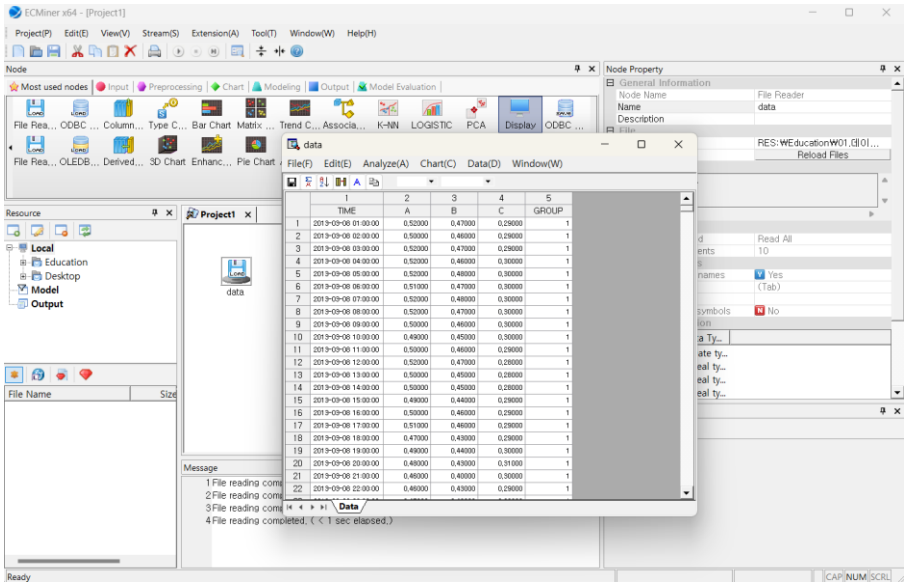
After reading the data for each node, the Data Browser can be started using one of the three methods below.

- Double click on the node
- Select Data Browser from the right-click context menu
- Click [View] – [Data Browser] on the toolbar

NOTE *In the case of Display Node, it is automatically executed when the project is performed. To run the project directly, you can double-click the Display Node in the output window of the Resource window after executing the project.*

Screen Layout

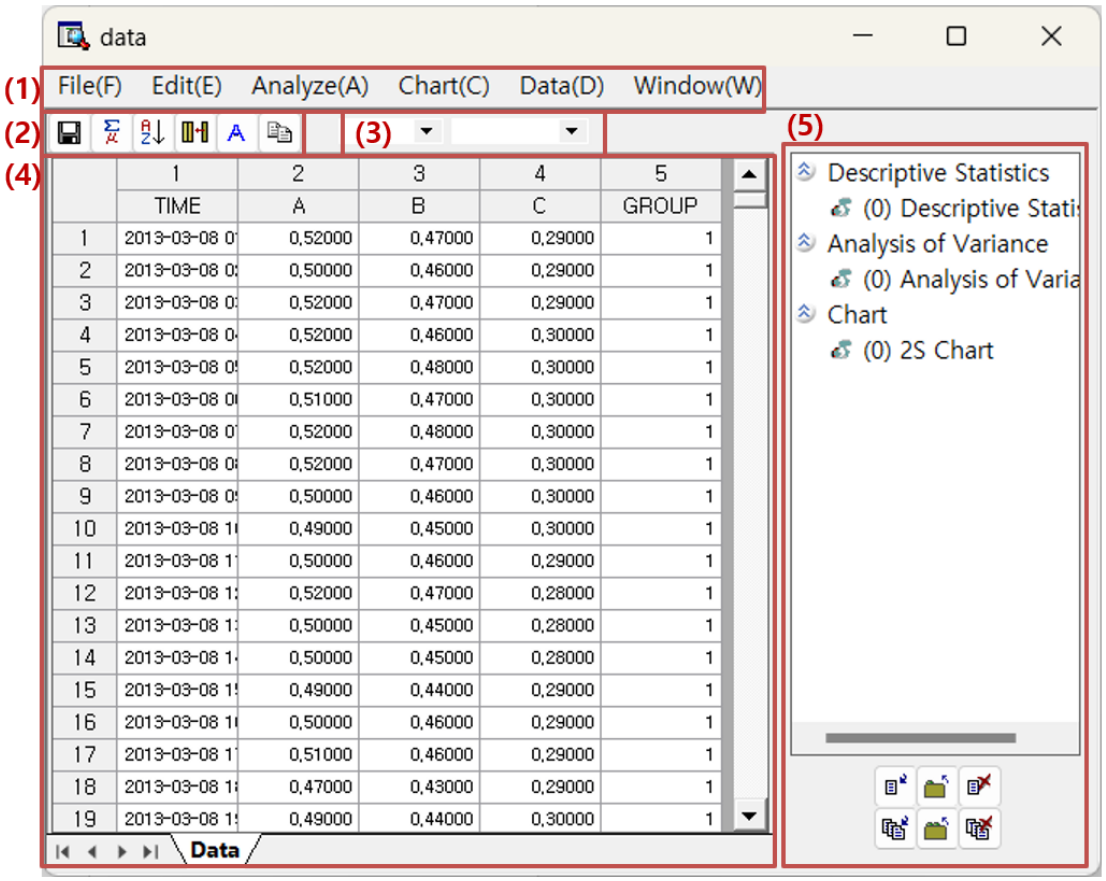
The Data Browser interface is as follows.



4.1 UI

4.1.1 Data Browser Screen Layout

The layout in Data Browser is as follows.



Windows Description

No.	Type of window	Function	Note
1	Menu	Menus in Data Browser	Additional description provided
2	Toolbar	Toolbar menus in Data Browser	Additional description provided
3	Variable Type	Specify and change data types and role	Data type: discrete, continuous Roles: independent variables, dependent variables
4	Viewer	Displays data or analysis results	Additional description provided

5	Results	Use the output management toolbar below to manage outputs like charts, statistical results, and variance analysis content.	Additional description provided

4.1.2 Main Menu

Menu

Menu Group	Name			Description
File	Reload			Reloads the data
	Save			Save data from Data Browser as a file.
Edit	Copy			Copy selected parts of rows, columns, and data areas.
	Find			Find data across the entire data.
	Find (In Field)			Find data in the data area for the selected field.
	Select All			Select the entire data.
Analyze	Data Summary Analysis	All Data		Summary statistics for all data
		Continuous Data		Summary statistics of continuous data
		Discrete Data		Frequency summary of discrete data
	Basic Statistics	Descriptive Statistics		Descriptive statistics for the selected variable
		Correlation Analysis	Display on Screen	Correlation values between continuous data
			Save to File	Save the correlation values
			Correlation Wheel	Correlations among continuous fields using a visual network.
		Mean Comparison Analysis	One-Sample t-Test	One sample t-test for the specified mean
			Two sample t-Test	Test for difference of mean for two groups
			Paired t-Test	Paired Samples t-test
		Proportion	One Sample	Test for one sample proportion

		Test	Proportion Test	whether it is significant from the hypothesized(specified) proportion
			Two Sample Proportion Test	Test for two independent groups whether it is significant difference between them
			One Sample Poisson Test	Test to determine if the observed rate of events (or counts) in a sample follows a specified Poisson distribution
			Two Sample Poisson Test	Test the rates of events between two independent samples to determine if they are significantly different. Counts of events follow a Poisson distribution
		Variance Test	One Sample Variance Test	Test whether the variance (or standard deviation) of a population is equal to a specified value
			Two Sample Variance Test	Test for two independent groups whether variances of two groups are significantly different.
		Normality Test		Test whether the data values in the selected field follow a normal distribution.
		Poisson Test		Test whether the data follows a Poisson distribution.
	Variance Analysis	One-Way ANOVA		One-way analysis of variance with one factor
		Two-Way ANOVA		Perform two-way analysis of variance with two factors
		General Linear Model		General form of ANOVA
	Regression Analysis	Multiple Linear Regression		Multiple linear regression through the linear equation of two or more independent variables.
		Orthogonal Regression		Regression to minimize the sum of squared distances from the observed points to the fitted line in

				all directions.
		Nonlinear Regression		Non-linear regression
	SPC	Process Capability Analysis		A process capability monitors process data within the desired specifications.
		Process Capability Result		Display whether the data is within the specified process
		Acceptance Sampling	Attributes Acceptance Sampling	Attribute sampling plan for acceptance and rejection in random sample taken from the batch.
			Quantitative Acceptance Sampling	Variable sampling plan for acceptance and rejection in random sample taken from the batch.
		Tolerance Intervals		Tolerance interval to satisfy the specified minimum proportion and confidence level
	Time Series Analysis	Time Series Models	Time Series Decomposition	Analyze and understand time series data through the decomposition of its integral components.
			Moving Average	Calculate the average of past data to smooth current data trends and forecast future values.
			Exponential Smoothing	Apply higher weights to recent data to forecast future values using a smoothing process.
			Trend Analysis	Examines data patterns over time to forecast future directions in trend analysis.
			ARIMA	Use to obtain the basic trend of time series data
			GARCH	Measures and predicts volatility.
			VAR	Models the relationships among a specific set of variables by considering multiple time series variables that influence one another
			ARMAX	Show time series data, fitted values,

				and the upper and lower limits of the fitted values.
		Time Series Test	Unit Root Test	Test whether a time series has a unit root.
			Granger Causality	Tests whether one time series is useful for predicting another time series.
			Cointegration Test	Test whether two time series are cointegrated.
			ARCH Test	Tests whether one time series can predict the time-varying volatility.
		Time Series Correlation	Cross-Correlation	Measure the similarity between two time series.
			Autocorrelation, Partial Autocorrelation	Correlation between the values of time series data and past data.
	Table	Frequency Table		Frequency of specific values for a field
		Cross Table		Frequency of values common to two variables for each field.
		Univariate Chi-Square Test		Test whether the collected data follows a multinomial distribution
		Independence Test		Test whether two categorical variables are independent.
	Probability Distribution	Parameter Estimation		Estimate the parameters that best fit the data and calculate the confidence interval of the estimated parameters based on the confidence level.
		Individual Distribution Identification		Calculate the Anderson-Darling statistic to test whether the given data conforms to a specific distribution.
	Nonparametric Test		One-Sample	Test whether the median of given data differs with a specified median.
			Independent Samples	Test whether the median between

			two groups is different.
		Paired Samples	Test whether the medians for each paired sample are different
		ANOVA - One-way	Test whether there are the differences among three or more medians.
	Accuracy Measurement	Classification	Evaluate the model's accuracy by comparing the predicted variables with the actual labels.
		Regression	Provide R-square, MAPE, MAD, and MSD using dependent and predictor variables.
	Gage R&R	Gage Run Chart	Shows the total observations obtained through the experiment.
		Gage Linearity and Bias Study	**Bias** is difference between the **average measured value** and the **true or reference value. T**he bias increases or decreases with the magnitude of the measurement, the gage is said to have poor linearity
		Gage R&&R Study (Crossed Design)	Evaluate the measurement system is reliable and that its variability is minimal compared to the overall process variability
		Gage R&&R Study (Nested Design)	Each part is measured only in a single measurement system.
Chart	2D Chart		Two-dimensional chart.
	3D Chart		Three-dimensional chart.
	Bar Chart		Bar chart.
	Box Chart		Box chart.
	Matrix Chart		Matrix chart.
	Pareto Chart		Pareto chart.
	Pie Chart		Pie chart.
	Multi-Chart		Multiple variables chart
Data	Sort		Sort data by the selected variable

	Derived Variables	Create a new derived variable
	Apply	Construct a stream of the manipulated preprocessing in the data browser
	Filter	Filter data by the specified variable option
	Freeze Interest Variables	Specify the variable of interest by selecting the column to freeze.
	Box-Cox Transformation	Transformation applied for non-normal data to nearly normal
	Johnson Transformation	Transformation applied for non-normal data to nearly normal
Window	Results	Manage the output results

4.1.3 Sub-menu (using mouse and keyboard)

The submenu in the Data Browser pops up right-clicking.

Delete Column

- **How to run**: Right-click on the column number or field name in the data area column header, and a menu will be shown.
- **Result**: The selected column will be deleted from the data browser. After executing 'Apply' following the column deletion, a filter node will be added to the stream, removing the selected column.

Delete Row

- **How to run**: Right-click on the row number in the row header of the data area and a menu will be shown.
- **Result**: The selected row will be deleted from the data browser. After executing 'Apply' following the row deletion, a Row Select Node will be added to the stream, removing the selected row

Hide Column

- **How to run**: Right-click on the column number or field name in the data area column header and a menu will be shown.
- **Result**: The selected column will be hidden in the Data Browser.

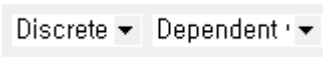
Cancel Hide Column

- **How to run**: Right-click on the column number or field name in the data area column header to show the hidden column number and field name. etc.
- **[Copy]/[Copy with column names]/[Select all]/[Export Data to Excel]** functions are additionally provided.

4.1.4 Toolbar

Toolbar in the Data Browser

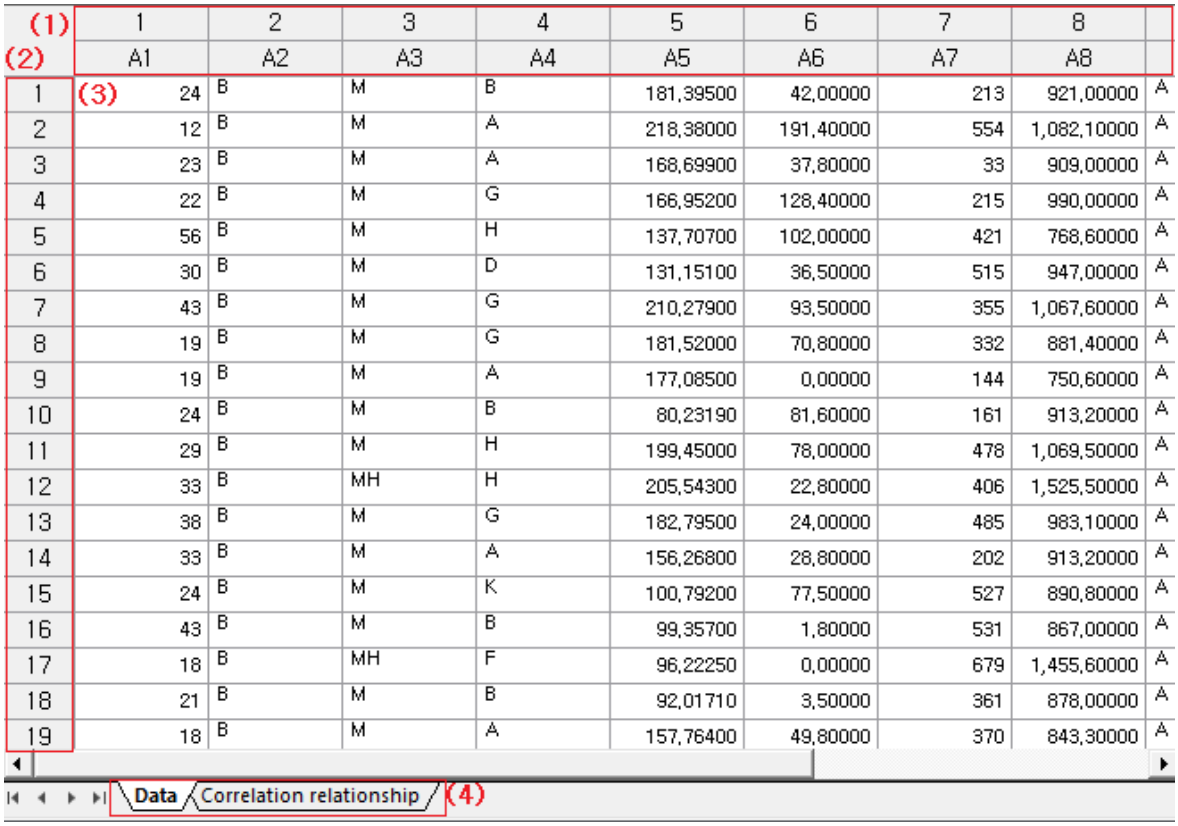
Toolbar Menu

Menu	Icon	Description	Note
Save		Save the current data to a file.	
Descriptive Statistics		Calculates descriptive statistics for the specified variable.	
Sort		Sorts the data for the specified variable.	
Derived Variable		Create derived variables.	
Chart		Select and configure the chart to be displayed in the data explorer.	
Apply		Apply the actions performed in the data explorer to the stream configuration.	
Result		Show or hide the results management window.	
Variable Type		Display and set the variable type.	

4.1.5 Data Area

You can explore data and edit rows and columns in the original datasheet.

Window Layout



1	24	B	M	B	181,39500	42,00000	213	921,00000
2	12	B	M	A	218,38000	191,40000	554	1,082,10000
3	23	B	M	A	168,69900	37,80000	33	909,00000
4	22	B	M	G	166,95200	128,40000	215	990,00000
5	56	B	M	H	137,70700	102,00000	421	768,60000
6	30	B	M	D	131,15100	36,50000	515	947,00000
7	43	B	M	G	210,27900	93,50000	355	1,067,60000
8	19	B	M	G	181,52000	70,80000	332	881,40000
9	19	B	M	A	177,08500	0,00000	144	750,60000
10	24	B	M	B	80,23190	81,60000	161	913,20000
11	29	B	M	H	199,45000	78,00000	478	1,069,50000
12	33	B	MH	H	205,54300	22,80000	406	1,525,50000
13	38	B	M	G	182,79500	24,00000	485	983,10000
14	33	B	M	A	156,26800	28,80000	202	913,20000
15	24	B	M	K	100,79200	77,50000	527	890,80000
16	43	B	M	B	99,35700	1,80000	531	867,00000
17	18	B	MH	F	96,22250	0,00000	679	1,455,60000
18	21	B	M	B	92,01710	3,50000	361	878,00000
19	18	B	M	A	157,76400	49,80000	370	843,30000

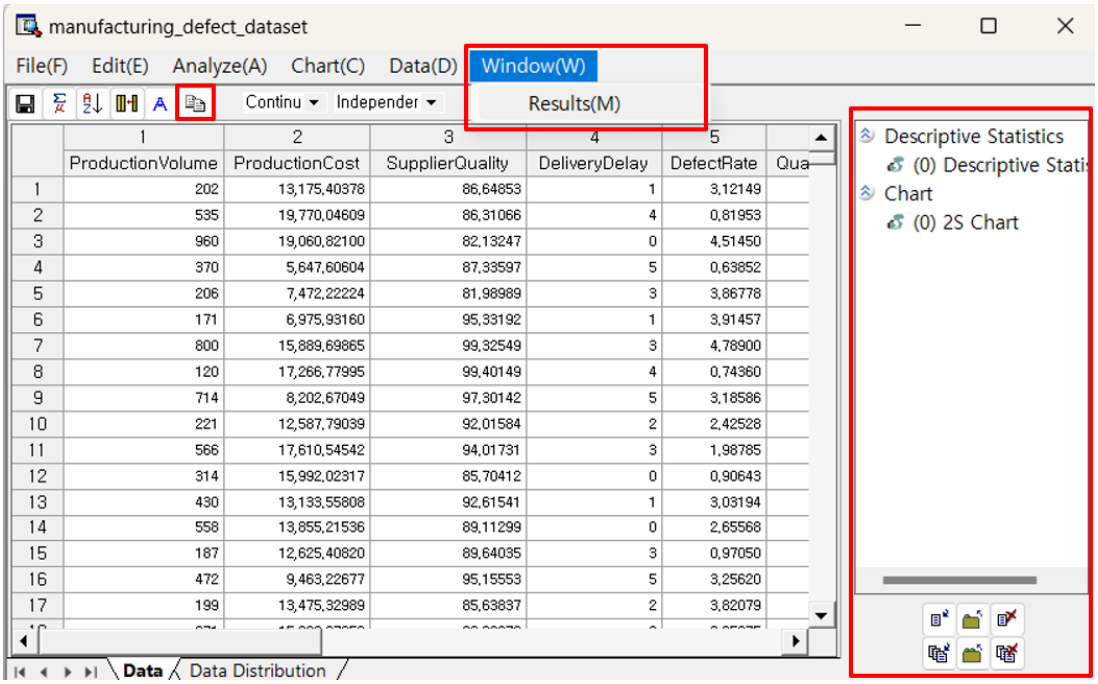
Window Description

No.	Type of window	Function	Note
1	Column header	**Column numbers** and **names i**n the data table. Right-click options: **Delete Column, Hide Column, Copy, Copy with column names, Select All,** and **Send Data to Excel**. Double-click the column border to auto-adjust the width.	
2	Row header	**Row numbers** and **names** in the data table. Right-click options: **Delete Row, Copy, Copy with column names, Select All,** and **Send Data to Excel.**	
3	Data Area	**Copy**, **Copy with column names**, **Delete**, **Select All**, and **Send Data to Excel**.	
4	Sheet tab	This tab controls the data. It adds sheets such as Data Distribution and Correlation.	

4.1.6 Results

Manage the results in the Data Browser by clicking [Window] - [Results] or the icon on toolbar. The results are grouped by type into categories such as “Chart” and “Descriptive Statistics”.

Window Layout



Results Toolbar Menu

Menu	Icon	Description	Etc.	Note
Show		Show the selected results window.		
Show all		Show all results windows.		
Close		Close the selected results window.		
Close all		Close all results windows.		
Remove		Remove the selected results window.		
Remove all		Remove all results windows.		

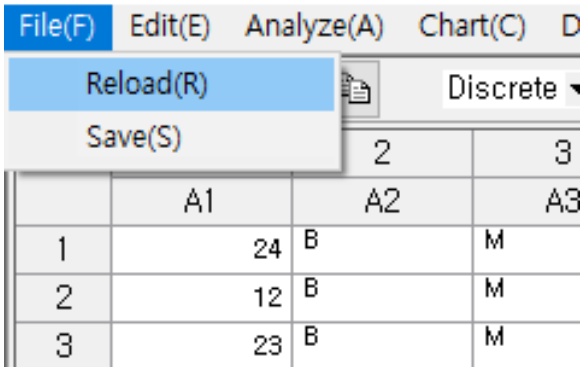
4.2 File

4.2.1 Reload

Reloads the data from a file or DB.

How to run

[File(F)] - [Reload(R)] in the menu.

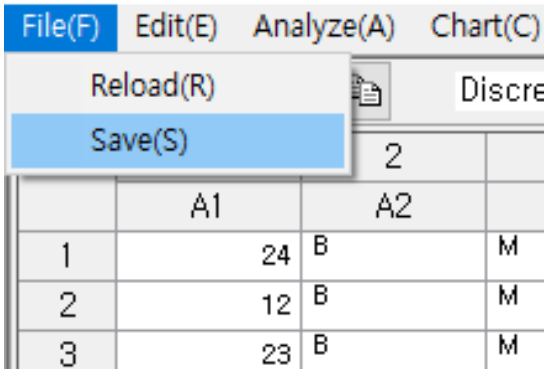


4.2.2 Save

Save the data.

How to run

[File(F)] - [Save(S)] in the menu or the  icon on the toolbar.



Save file

Files can be saved with extensions such as "txt", "csv", and “ecl".

4.3 Analyze

4.3.1 Data Summary Statistics

---

(1) All Data

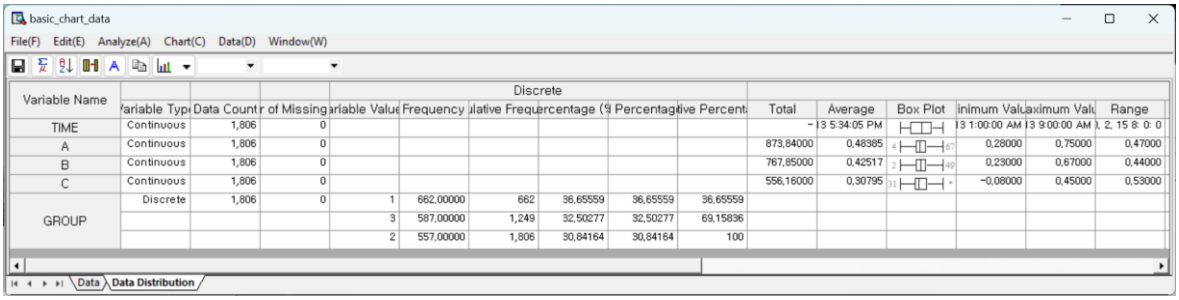
Data Summary Statistics – All Data displays summary statistics of all the variables.

How to run

[Analyze] – [Data Summary Statistics] – [All Data]

	1	2	3	4	5	6	7	8	9	10
	A1	A2	A3	A4	A5	A6	A7	A8	A9	customer_churn
1	24	B	M	B	181,39500	42,00000	213	921,00000	A	0
2	12	B	M	A	218,38000	191,40000	554	1,082,10000	A	1
3	23	B	M	A	168,69900	37,80000	33	909,00000	A	1
4	22	B	M	G	166,95200	128,40000	215	990,00000	A	0
5	56	B	M	H	137,70700	102,00000	421	768,60000	A	1
6	30	B	M	D	131,15100	36,50000	515	947,00000	A	1
7	43	B	M	G	210,27900	93,50000	355	1,067,60000	A	1
8	19	B	M	G	181,52000	70,80000	332	881,40000	A	0
9	19	B	M	A	177,08500	0,00000	144	750,60000	A	1
10	24	B	M	B	80,23190	81,60000	161	913,20000	A	1
11	29	B	M	H	199,45000	78,00000	478	1,069,50000	A	1
12	33	B	MH	H	205,54300	22,80000	406	1,525,50000	A	1
13	38	B	M	G	182,79500	24,00000	485	983,10000	A	1
14	33	B	M	A	156,26800	28,80000	202	913,20000	A	1
15	24	B	M	K	100,79200	77,50000	527	890,80000	A	1
16	43	B	M	B	99,35700	1,80000	531	867,00000	A	0
17	18	B	MH	F	96,22250	0,00000	679	1,455,60000	A	1
18	21	B	M	B	92,01710	3,50000	361	878,00000	A	0

Results



Variable Name	Variable Type	Data Count	Number of Missing	Discrete Variable Value	Frequency	Relative Frequency	Percentage (%)	Percentage Percent	Total	Average	Box Plot	Minimum Value	Maximum Value	Range
TIME	Continuous	1,806	0						-13 5:34:05 PM			13 1:00:00 AM	13 9:00:00 AM	0. 2. 15 8: 0: 0
A	Continuous	1,806	0						873,84000	0,48385		0,28000	0,75000	0,47000
B	Continuous	1,806	0						767,85000	0,42517		0,23000	0,67000	0,44000
C	Continuous	1,806	0						556,16000	0,30795		-0,08000	0,45000	0,53000
GROUP	Discrete	1,806	0	1	662,00000	662	36,65559	36,65559	36,65559					
				3	587,00000	1,249	32,50277	32,50277	69,15836					
				2	557,00000	1,806	30,84164	30,84164	100					

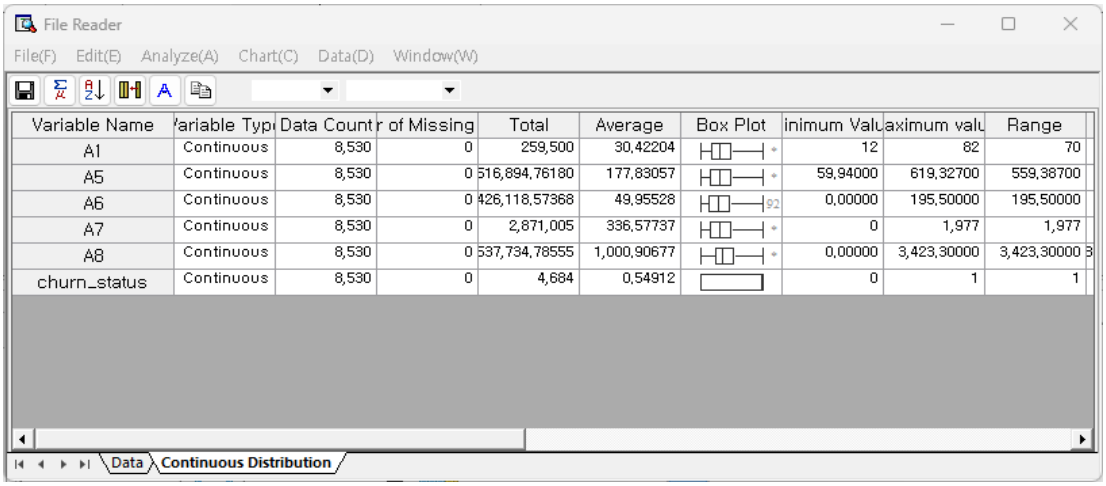
(2) Continuous Data

Data Summary Statistics – Continuous Data shows descriptive statistics with box plot for all continuous data.

How to run

[Analyze] – [Data Summary Statistics] – [Continuous Data]

Results



Output Statistics

Total, Average, Box plot, Number of Missing Values, Minimum Value, Maximum Value, Range, Variance, Standard Deviation, Kurtosis, Skewness, Median, and Quartile are supported.

NOTE *The numbers at either end of the box plot indicates the number of outliers. If the number of outliers exceeds 100, it is displayed as '*'.*

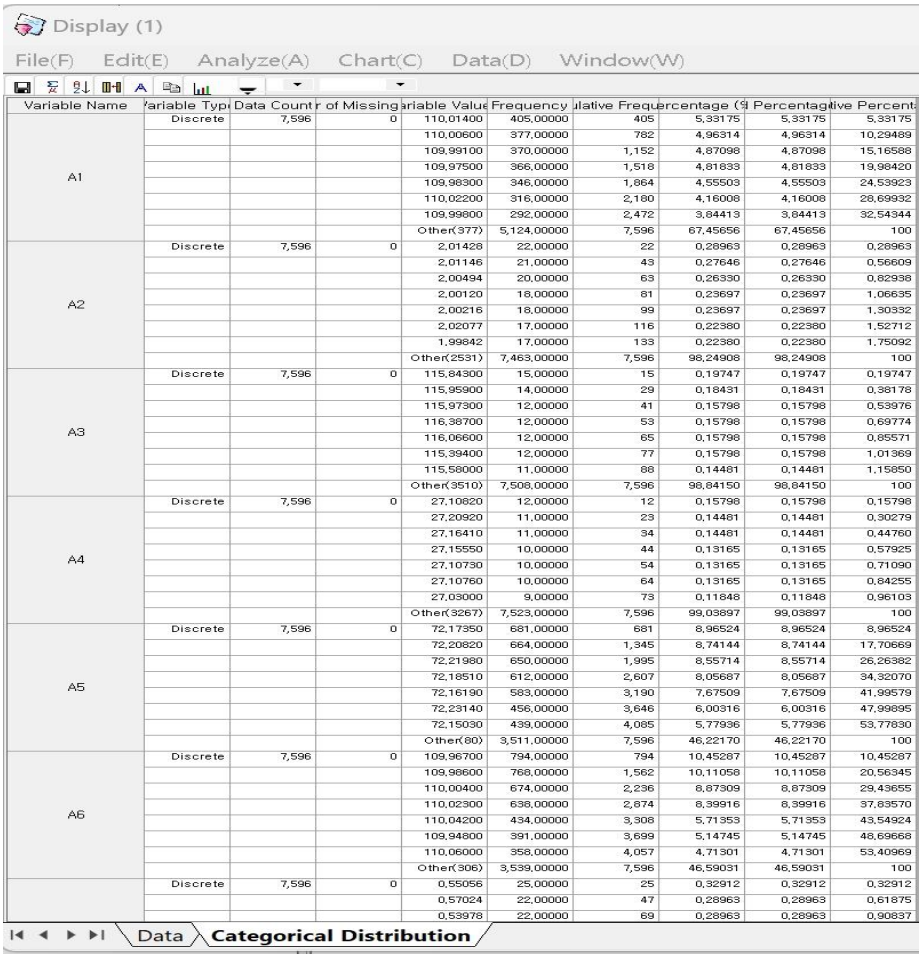
(3) Discrete Data

Data Summary Statistics – Discrete Data shows descriptive statistics for discrete data.

How to run

[Analyze] – [Data Summary Statistics] – [Discrete Data]

Results


Display (1)

File(F) Edit(E) Analyze(A) Chart(C) Data(D) Window(W)

Variable Name	Variable Typ	Data Count	r of Missing	ariable Value	Frequency	ulative Freque	ercentage (%	Percentage	tive Percent
A1	Discrete	7,596	0	110,01400	405,00000	405	5,33175	5,33175	5,33175
				110,00600	377,00000	782	4,96314	4,96314	10,29489
				109,99100	370,00000	1,152	4,87098	4,87098	15,16588
				109,97500	366,00000	1,518	4,81833	4,81833	19,98420
				109,98300	346,00000	1,864	4,55503	4,55503	24,53923
				110,02200	316,00000	2,180	4,16008	4,16008	28,69932
				109,99800	292,00000	2,472	3,84413	3,84413	32,54344
				Other(377)	5,124,00000	7,596	67,45656	67,45656	100
A2	Discrete	7,596	0	2,01428	22,00000	22	0,28963	0,28963	0,28963
				2,01146	21,00000	43	0,27646	0,27646	0,56609
				2,00494	20,00000	63	0,26330	0,26330	0,82938
				2,00120	18,00000	81	0,23697	0,23697	1,06635
				2,00216	18,00000	99	0,23697	0,23697	1,30332
				2,02077	17,00000	116	0,22380	0,22380	1,52712
				1,99842	17,00000	133	0,22380	0,22380	1,75092
				Other(2531)	7,463,00000	7,596	98,24908	98,24908	100
A3	Discrete	7,596	0	115,84300	15,00000	15	0,19747	0,19747	0,19747
				115,95900	14,00000	29	0,18431	0,18431	0,38178
				115,97300	12,00000	41	0,15798	0,15798	0,53976
				116,38700	12,00000	53	0,15798	0,15798	0,69774
				116,06600	12,00000	65	0,15798	0,15798	0,85571
				115,39400	12,00000	77	0,15798	0,15798	1,01369
				115,58000	11,00000	88	0,14481	0,14481	1,15850
				Other(3510)	7,508,00000	7,596	98,84150	98,84150	100
A4	Discrete	7,596	0	27,10820	12,00000	12	0,15798	0,15798	0,15798
				27,20920	11,00000	23	0,14481	0,14481	0,30279
				27,16410	11,00000	34	0,14481	0,14481	0,44760
				27,15550	10,00000	44	0,13165	0,13165	0,57925
				27,10730	10,00000	54	0,13165	0,13165	0,71090
				27,10760	10,00000	64	0,13165	0,13165	0,84255
				27,03000	9,00000	73	0,11848	0,11848	0,96103
				Other(3267)	7,523,00000	7,596	99,03897	99,03897	100
A5	Discrete	7,596	0	72,17350	681,00000	681	8,96524	8,96524	8,96524
				72,20820	664,00000	1,345	8,74144	8,74144	17,70669
				72,21980	650,00000	1,995	8,55714	8,55714	26,26382
				72,18510	612,00000	2,607	8,05687	8,05687	34,32070
				72,16190	583,00000	3,190	7,67509	7,67509	41,99579
				72,23140	456,00000	3,646	6,00316	6,00316	47,99895
				72,15030	439,00000	4,085	5,77936	5,77936	53,77830
				Other(80)	3,511,00000	7,596	46,22170	46,22170	100
A6	Discrete	7,596	0	109,96700	794,00000	794	10,45287	10,45287	10,45287
				109,98600	768,00000	1,562	10,11058	10,11058	20,56345
				110,00400	674,00000	2,236	8,87309	8,87309	29,43655
				110,02300	638,00000	2,874	8,39916	8,39916	37,83570
				110,04200	434,00000	3,308	5,71353	5,71353	43,54924
				109,94800	391,00000	3,699	5,14745	5,14745	48,69668
				110,06000	358,00000	4,057	4,71301	4,71301	53,40969
				Other(306)	3,539,00000	7,596	46,59031	46,59031	100
	Discrete	7,596	0	0,55056	25,00000	25	0,32912	0,32912	0,32912
				0,57024	22,00000	47	0,28963	0,28963	0,61875
				0,53978	22,00000	69	0,28963	0,28963	0,90837

Data | Categorical Distribution

Output Statistics

Number of Missing Values, Variable Values, Frequency, Cumulative Frequency, Percentage, Valid Percentage, and Cumulative Percentage.

4.3.2 Basic Statistics

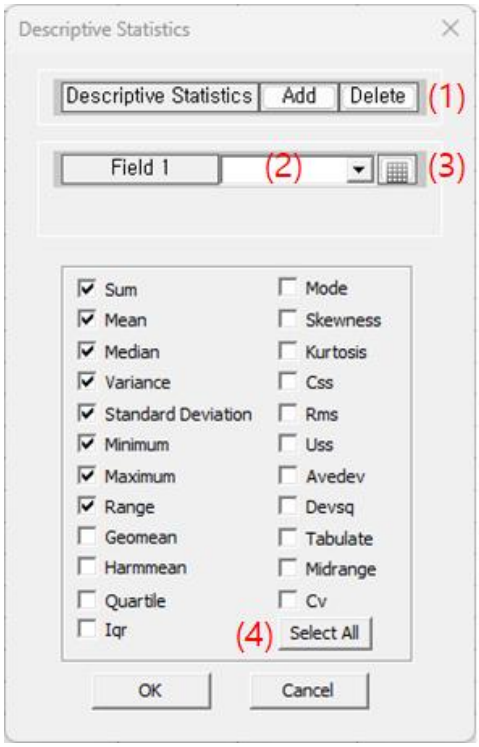
4.3.2.1 Descriptive Statistics

Basic Statistics – Descriptive Statistics shows descriptive statistics of the selected variables.

How to run

[Analyze] – [Basic Statistics] – [Descriptive Statistics]

Descriptive Statistics options



- Add a variable or remove a variable using Add and Delete button.
- Select the variable for the descriptive statistics.
- Button is another way to select the column header in the data browser
- Click the boxes to select or deselect the statistics.

Results

Sum	1516894,7618
Mean	177,83057
Min	59,94
Max	619,327
Range	559,387
Variance	6866,26352
Standard Deviation	82,86292
Kurtosis	1,61591
Skewness	1,09217
Median	165,1835
Quartile Q1	113,913
Quartile Q2	165,1835
Quartile Q3	222,191

Output Statistics

The chosen statistics are shown. For all of the statistics, Sum, Mean, Median, Variance, Standard Deviation, Minimum, Maximum, Range, Geamean, Harmmean, Quartile, Iqr, Mode, Skewness, Kurtosis, Css, Rms, Uss, Avedev, Devsq, Tabulate, Midrange, and Cv are supported.

4.3.2.2 Correlation Analysis

---

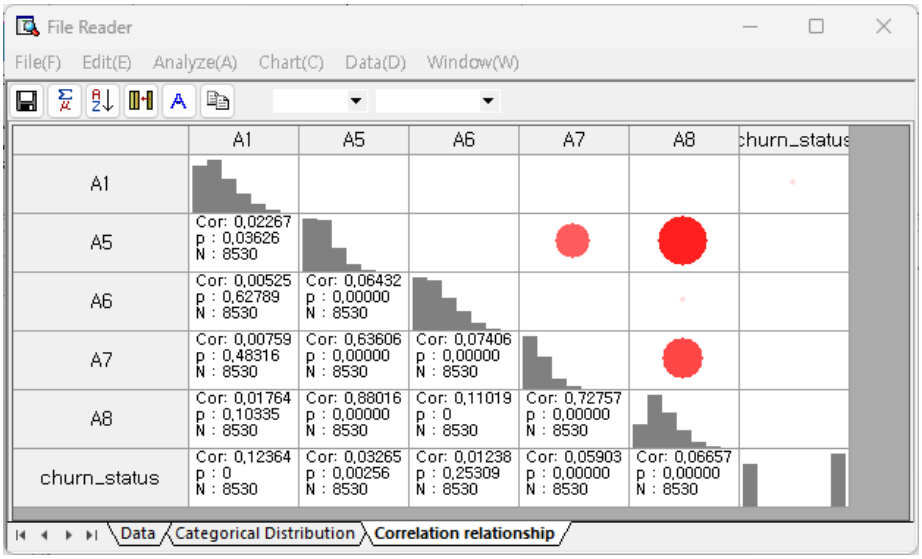
(1) Display on Screen

Correlation Analysis – Display on Screen shows the correlation.

How to run

[Analyze] – [Basic Statistics] – [Correlation Analysis] – [Display on Screen]

Results



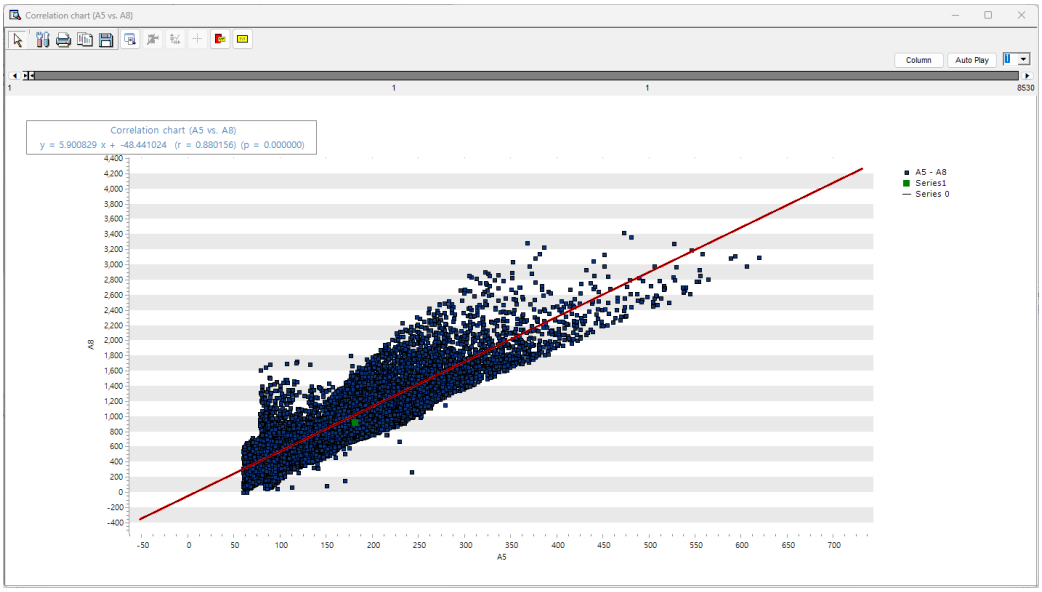
Output Statistics

Cor (correlation value), and p (p-value)

Correlation chart

Distribution chart with a trend line is displayed double-clicking the selected correlation

table.

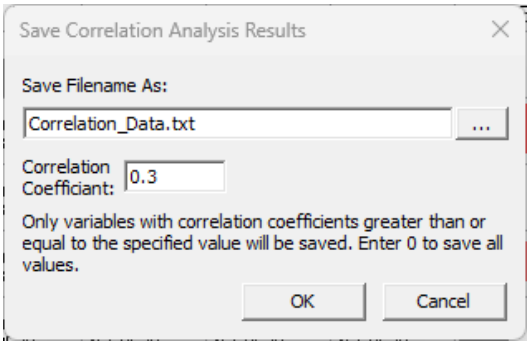


(2) Save to File

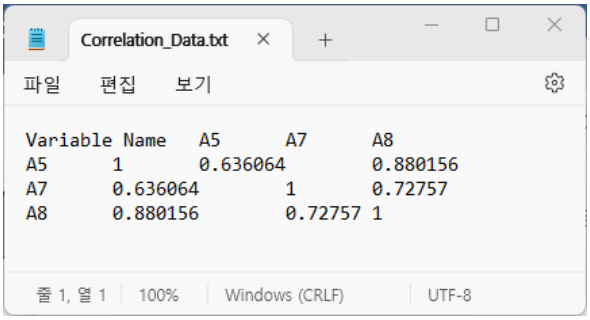
Correlation Analysis – Save to File do save the correlation table as text file.

How to run

[Analyze] – [Basic Statistics] – [Correlation Analysis] – [Save on File]. Specify the file name, directory, and the correlation coefficient value. Results are stored only for variables that have correlation ≥ the specified value.



Results



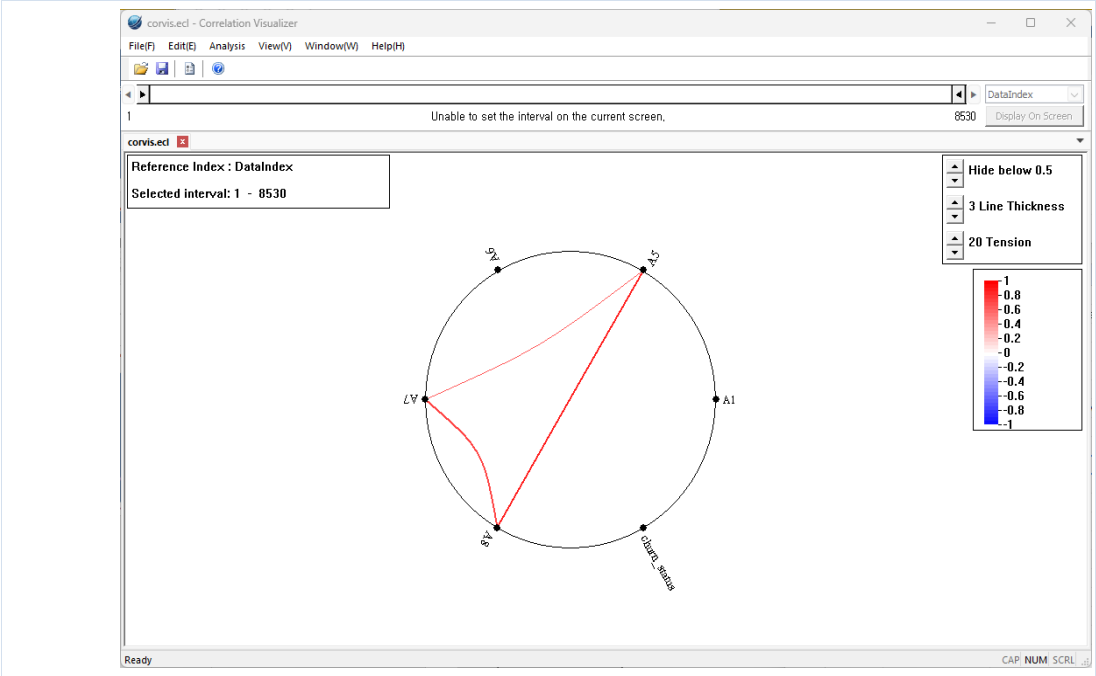
(3) Correlation Wheel

Correlation Analysis – Correlation Wheel shows the correlation between variables using the color and thickness of the connecting lines.

How to run

[Analyze] – [Basic Statistics] – [Correlation Analysis] – [Correlation Wheel]

Results



Red means positive correlation, blue means negative correlation, and the thicker the connecting line, the larger the absolute value of the correlation coefficient. On the top right corner, there is a menu to adjust the coefficient threshold value, line thickness, and tension. The coefficient threshold allows you to selectively view the meaningful correlation between variables.

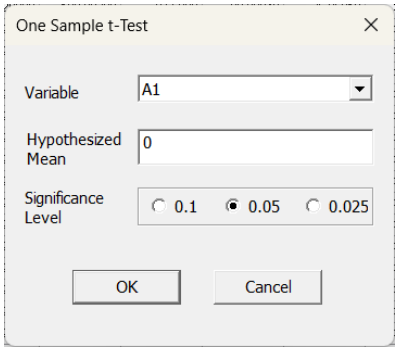
4.3.2.3 Mean Comparison Analysis

(1) One-Sample t-test

One-sample t-test is to test whether the mean of one sample differs significantly from the specified mean.

How to run

[Analyze]-[Basic Statics]–[Mean Comparison Analysis]]-[One Sample t-Test]



Select a **variable**, and define the **hypothesized mean**, and **confidence level**.

Result

One-sample t-test statistics result.

One Sample t-Test

One-Sample Statistics

Variable Name	Data Count	Average	Standard Deviation	Mean Standard Error
A1	7596	109.35402	1.74303	0.02000

One-Sample Test

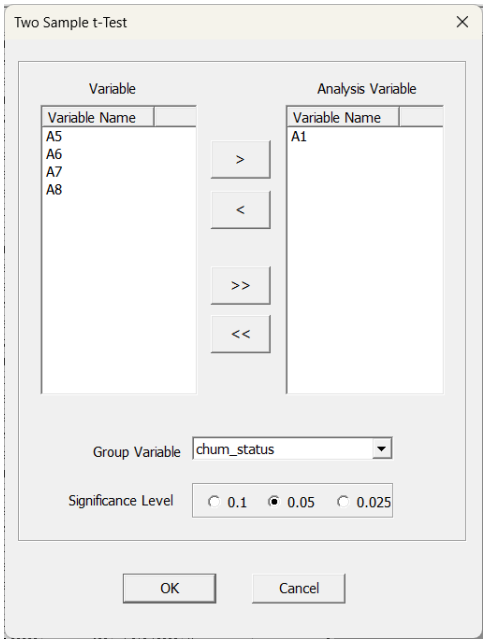
	t	Freedom Degree	p-Value(Two-sided)	Average Difference	95% confidence interval of the difference
A1	5467.91159	7595	0.00000	109.35402	(109.315, 109.393)

(2) Two Sample t-Test

Two Sample t-Test is to compare the average of two independent samples.

How to run

[Analyze]-[Basic Statics]–[Mean Comparison Analysis]-[Two Sample t-Test]



Select **analysis variables** and **Group variable**.

Result

Two-sample t-test statistics result.

Two Sample t-Test

Group statistics

	chum_status	Number of Data Points	Average	Standard Deviation	Standard Error of the Mean
A1	0	3846	28,67447	12,59226	0,20305
	1	4684	31,85696	12,80715	0,18713

Two Sample Test

Field Name	Assumption	F	Significant Probability (Levene)	t	Freedom Degree	p-Value (Two-Sided)	Average Difference	Standard error of the difference	95% confidence interval of the difference
A1	Equal Variance Assumption	0,08146	0,77534	-2679,89508	8528	0,00000	-3,18249	0,00119	(-3,185, -3,180)
	Not assuming equal variance			-11,52546	8258,64568	0,00000	-3,18249	0,27613	(-3,724, -2,641)

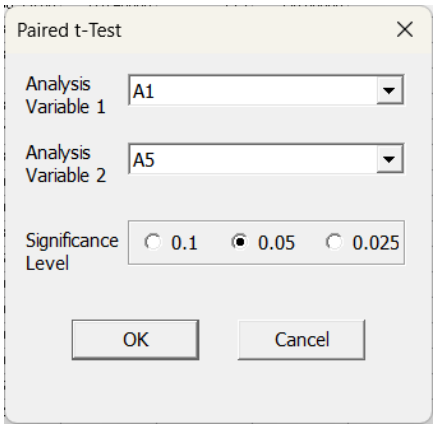
---

(3) Paired Samples

Paired t-test is to test whether the mean difference between two paired observations. Measurements are taken on the same subjects before and after a treatment, or paired observation

How to run

[Analyze]-[Basic Statics]–[Mean Comparison Analysis]-[Paired t-Test]



Result

Paired t-Test results.

T-Test(Paired Sample)

Paired Sample Statistics

	Average	Number of Data Points	Standard Deviation	Standard Error of the Mean
A1	30,42204	8530	12,80825	0,13868
A5	177,83057	8530	82,86292	0,89719

Paired Sample Test

	Average	Standard Deviation	Standard Error of the Mean	95% confidence interval of the difference	t	Freedom Degree	Significant Probability (Two-Sided)
A1 - A5	-147,40853	83,55948	0,90473	(-149,182, -145,635)	-162,93013	8529	0,00000

4.3.2.4 Proportion Test

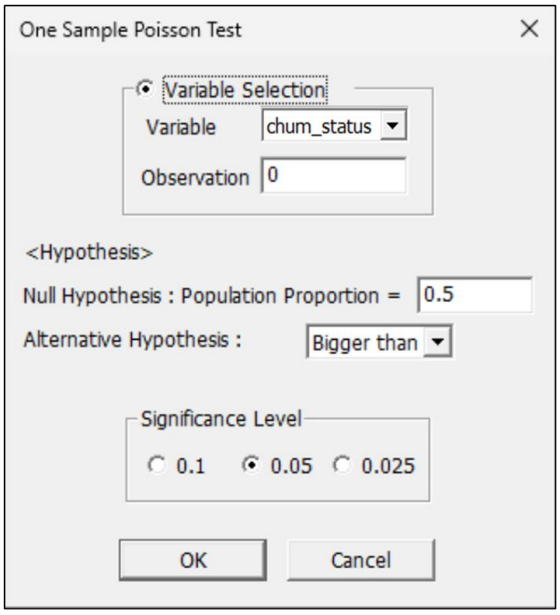
---

(1) One sample Proportion Test

One sample proportion test of the specified ratio with confidence interval.

How to run

[Analyze] - [Basic Statistics] - [Proportion Test] - [One sample Proportion Test]



- Select a variable. Choose event class among the class of variable.
- Set hypothesis parameter
- Set confidence interval.

Results

The smaller a P-value is, the greater the possibility of rejecting a null hypothesis is.

One Sample Proportion Test

Hypothesis Testing and Confidence Interval

Null Hypothesis : Population Proportion = 0.500000, Alternative Hypothesis : Bigger than

Variable	Number of Observations	Number of Events	Sample ratio	95% C.I	z-value	p-value
chum_status	8530	4684	0.54912	(,)	9.07339	0

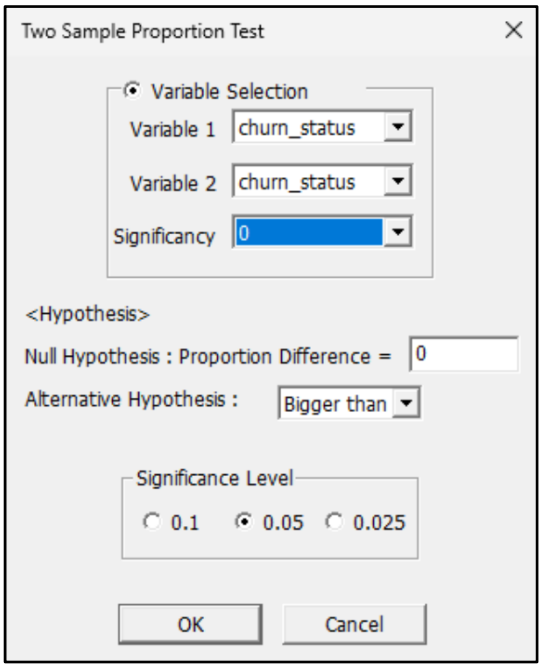
---

(2) Two sample Proportion Test

Two sample proportion test is to test significant difference between two groups.

How to run

[Analyze] - [Basic Statistics] - [Proportion Test] - [Two Sample Proportion Test]



- Select variables. Choose event class among the class of variable.
- Set hypothesis parameter.
- Set confidence level

Results

The smaller a P-value is, the greater the possibility of rejecting a null hypothesis is. Therefore, the proportion difference between two populations can be claimed like the type of an alternative hypothesis.

Two Proportion Test

Hypothesis Testing and Confidence Interval

Null Hypothesis : Population Proportion Difference = 0.000000, Alternative Hypothesis : Bigger than

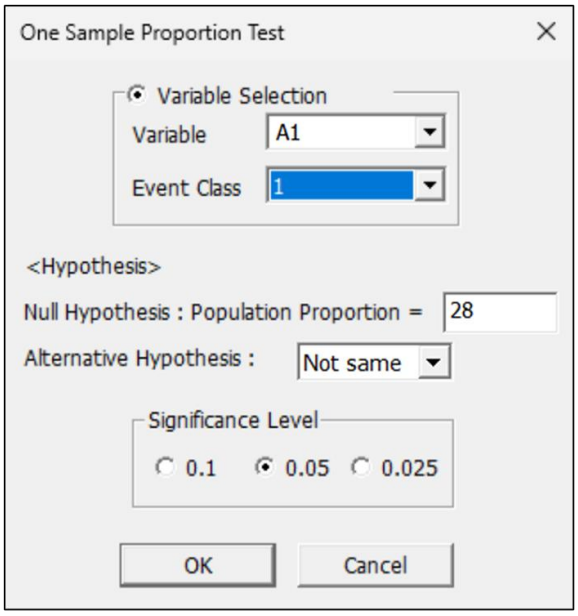
Variable	Number of Observations	Number of Events	Sample ratio	95% C.I	z-value	p-value
A2	2763	1726	0.62468	(,)	-38.95522	1
A9	2763	2737	0.99059	(,)	-38.95522	1

(3) One-Sample Poisson Test

Tests whether the event rate in a sample follows a Poisson distribution and matches the expected value.

How to run

[Analyze] - [Basic Statistics] – [Proportion Test] – [One Sample Poisson Test]



- The following procedures are performed after a desired method is chosen between variable selection method and manual input method.
- The variable selection method uses an integer variable and selects a corresponding variable and decides the length of an observation value.
- In the manual input method, the number of a trial and the number of an event occurrence is entered directly.
- For verification the population ratio is set in a null hypothesis, and the type and significance level of an alternative hypothesis is selected.

Results

The test results of a chosen method are presented in a table as follows; The smaller a P-value is, the greater the possibility of rejecting a null hypothesis is. Therefore, the population proportion can be claimed like the type of an alternative hypothesis.

One Sample Poisson Test

Hypothesis Testing and Confidence Interval

Null Hypothesis : Population Proportion = 28.000000, Alternative Hypothesis : Not same

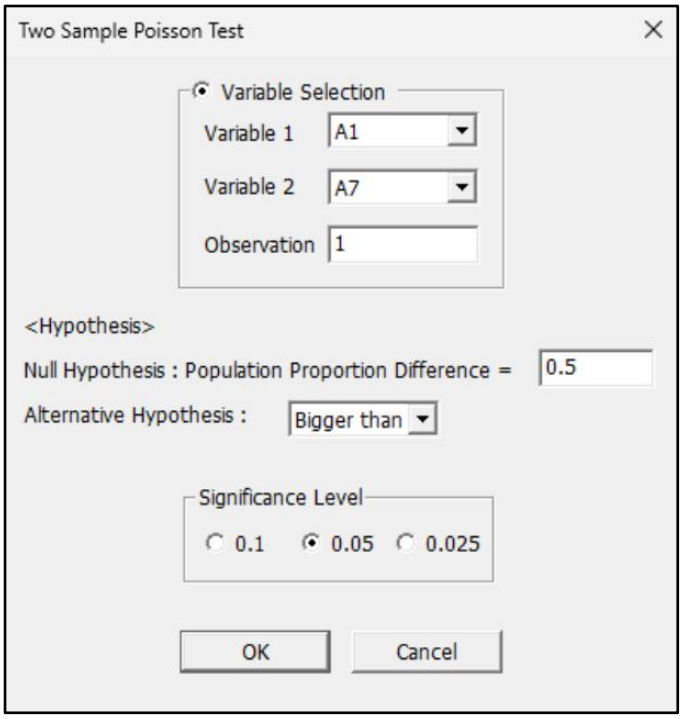
Variable	Number of samples	Number of Occurrences	Observation Length	Incidence Rate	Average Occurrence Count	95% C.I	z-value	p-value
A1	8530	259500	1	30.42204	30.42204	(,)	42.27434	0

(4) Two Sample Poisson Test

Tests whether the event rates of two independent samples, assumed to follow Poisson distributions, are significantly different.

How to run

[Analyze] - [Basic Statistics] – [Proportion Test] – [Two Sample Poisson Test],



- The following procedures are performed after a desired method is chosen between variable selection method and manual input method.
- The variable selection method uses an integer variable and selects corresponding two variables and decides the length of an observation value.
- In the manual input method, the number of two trials and the number of an event occurrence is entered directly.
- For verification the population ratio is set in a null hypothesis, and the type and significance level of an alternative hypothesis is selected.

Results

The test results of a chosen method are presented in a table as follows; The smaller a P-value is, the greater the possibility of rejecting a null hypothesis is. Therefore, the difference of the population proportion of 2-sample Poisson can be claimed like the type

of an alternative hypothesis.

Two Sample Poisson Test

Hypothesis Testing and Confidence Interval

Null Hypothesis : Population Proportion Difference = 0.500000, Alternative Hypothesis : Bigger than

Variable	Number of samples	Number of Occurrences	Incidence Rate	Average Occurrence Count	Rate Difference	95% C.I	z-value	p-value
A1	8530	259500	30.42204	30.42204	-306.15533	(,)	-1478.40136	1
A7	8530	2871005	336.57737	336.57737	-306.15533	(,)	-1478.40136	1

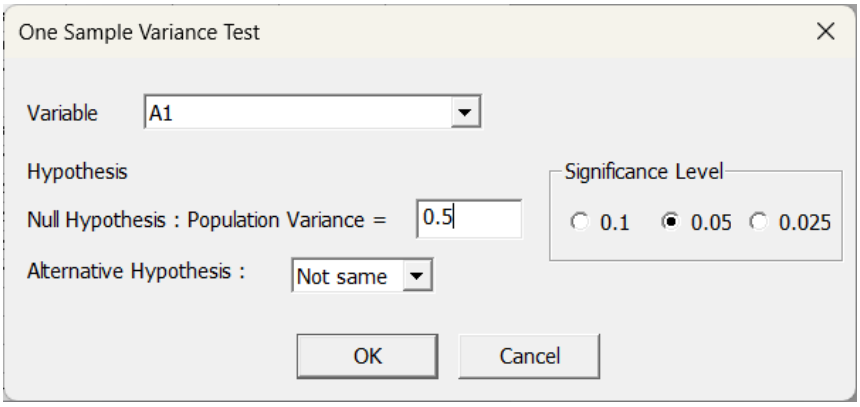
4.3.2.5 Variance Test

(1) One Sample Variance Test

Data explorer provides the function of single sample variance test that calculates the confidence interval and tests a hypothesis for the variance of a continuous variable.

How to run

[Analyze] - [Basic Statistics] - [Variance Test] - [One Sample Variance Test]



- A target variable is chosen for single sample variance test.
- For verification, the variance value of a population is set in a null hypothesis and the type and significance level of an alternative hypothesis is selected.

Results

The test results of a chosen method are presented in a table as follows; The smaller a P-value is, the greater the possibility of rejecting a null hypothesis is. Therefore, the variance of a population can be claimed like the type of an alternative hypothesis

Single Sample Variance Test

Hypothesis Testing and Confidence Interval

Null Hypothesis : Population Variance = 5.000000, Alternative Hypothesis : Not same

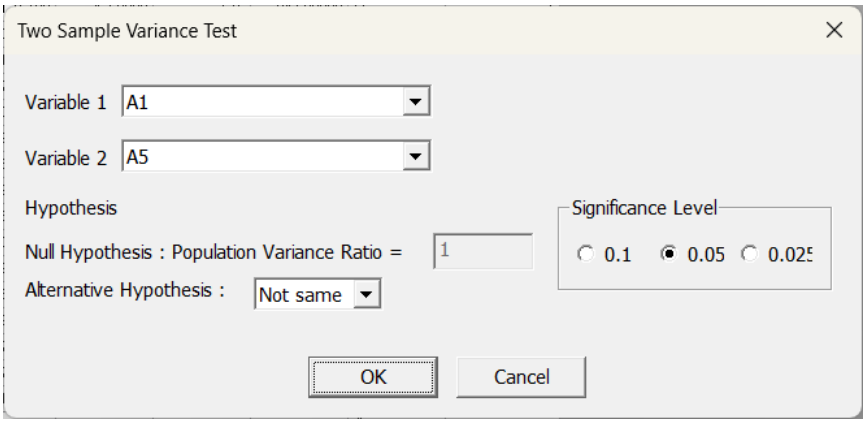
Variable	Number of Observations	Standard Deviation	Variance	95% C.I	Freedom Degree	x2 value	p-value
A1	8530	12.80825	164.05120	(159.236707 , 169.088444)	8529	279838.53130	0

(2) Two Sample Variance Test

Data explorer provides the function of double sample variance test that calculates the confidence interval and tests a hypothesis for the ratio of the variance of two continuous variables.

How to run

[Analyze] - [Basic Statistics] - [Variance Test] - [Two Sample Variance Test]



- Two variables are chosen for double sample variance test.
- For verification, the variance ratio of a population is set in a null hypothesis and the type and significance level of an alternative hypothesis is selected.

Results

The test results of a chosen method are presented in a table as follows; The smaller a P-

value is, the greater the possibility of rejecting a null hypothesis is. Therefore, the ratio of the variance of two populations can be claimed like the type of an alternative hypothesis.

Two Sample Variance Test

Hypothesis Testing and Confidence Interval

Null Hypothesis : Population Variance Ratio = 1.000000, Alternative Hypothesis : Bigger than

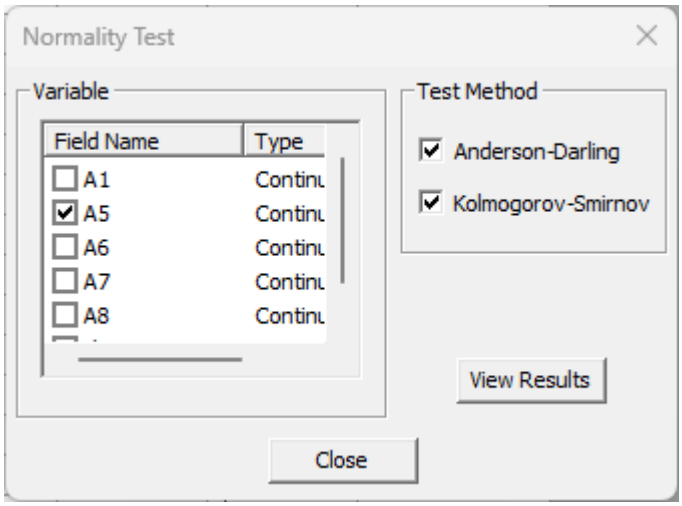
Variable	Number of Observations	Standard Deviation	Variance	Variance Ratio	95% C.I	Freedom Degree1	Freedom Degree2	f value	p-value
A1	8530	12,80825	164,05120	0,02389	(0,023056 , 0,000000)	8529	8529	0,02389	1
A5	8530	82,86292	6866,26352	0,02389	(0,023056 , 0,000000)	8529	8529	0,02389	1

4.3.2.6 Normality Test

Normality test is for testing whether data is normally distributed. It takes Anderson-Darling and Kolmogorov-Smirnov tests.

How to run

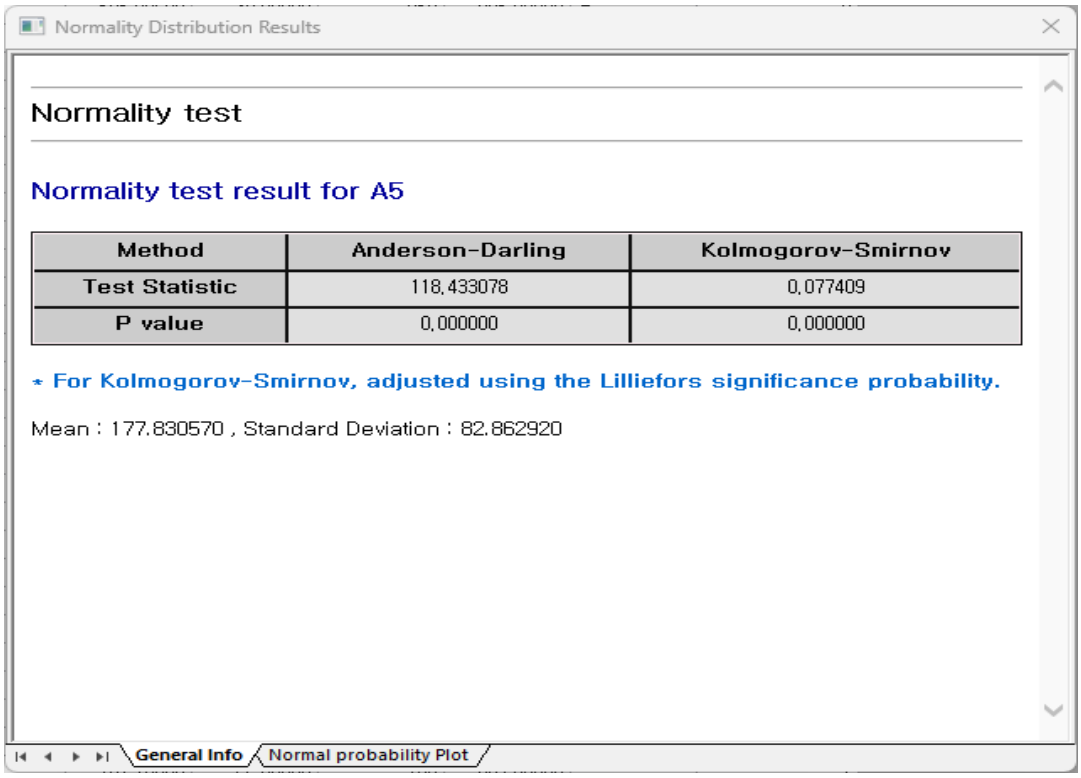
[Analyze] – [Basic Statistics] - [Normality Test]



Select a variable and choose the test method.

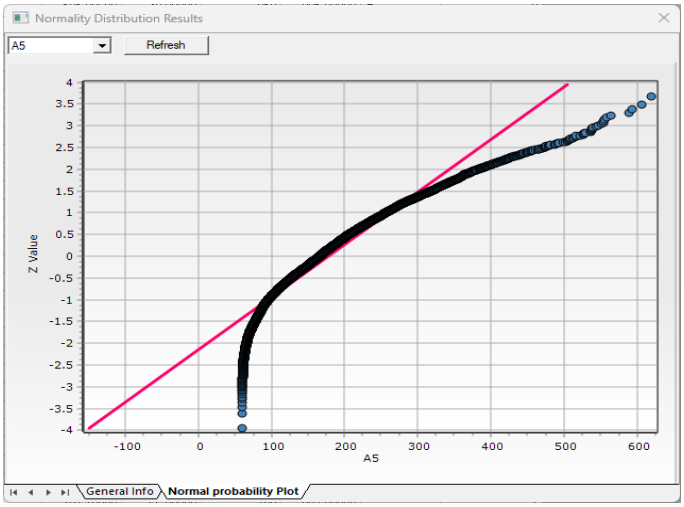
Results

The smaller a p-value is, the more normal distribution is not.



- **Normal probability plot**

  It shows whether data follows a normal distribution. If the data points form a roughly straight line, it suggests that the data is normally distributed.



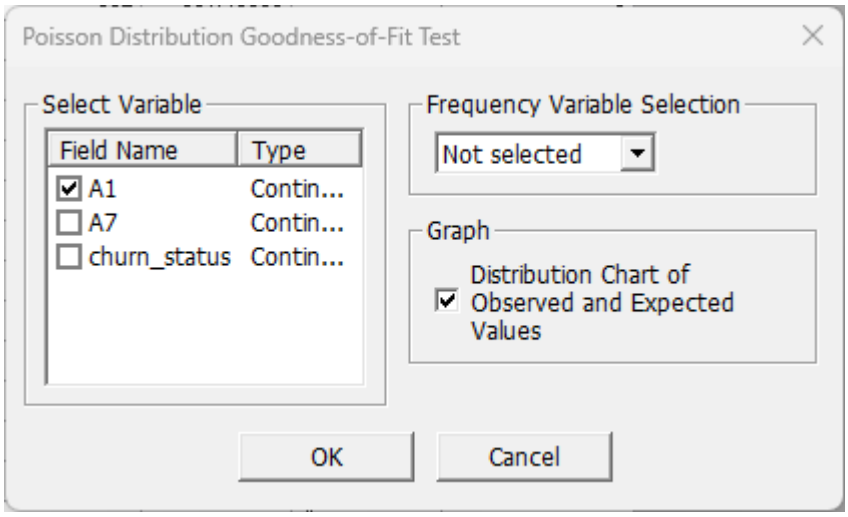
4.3.2.7 Poisson Test

The Poisson goodness-of-fit test is a technique that tests whether the collected data follows a Poisson distribution.

How to run

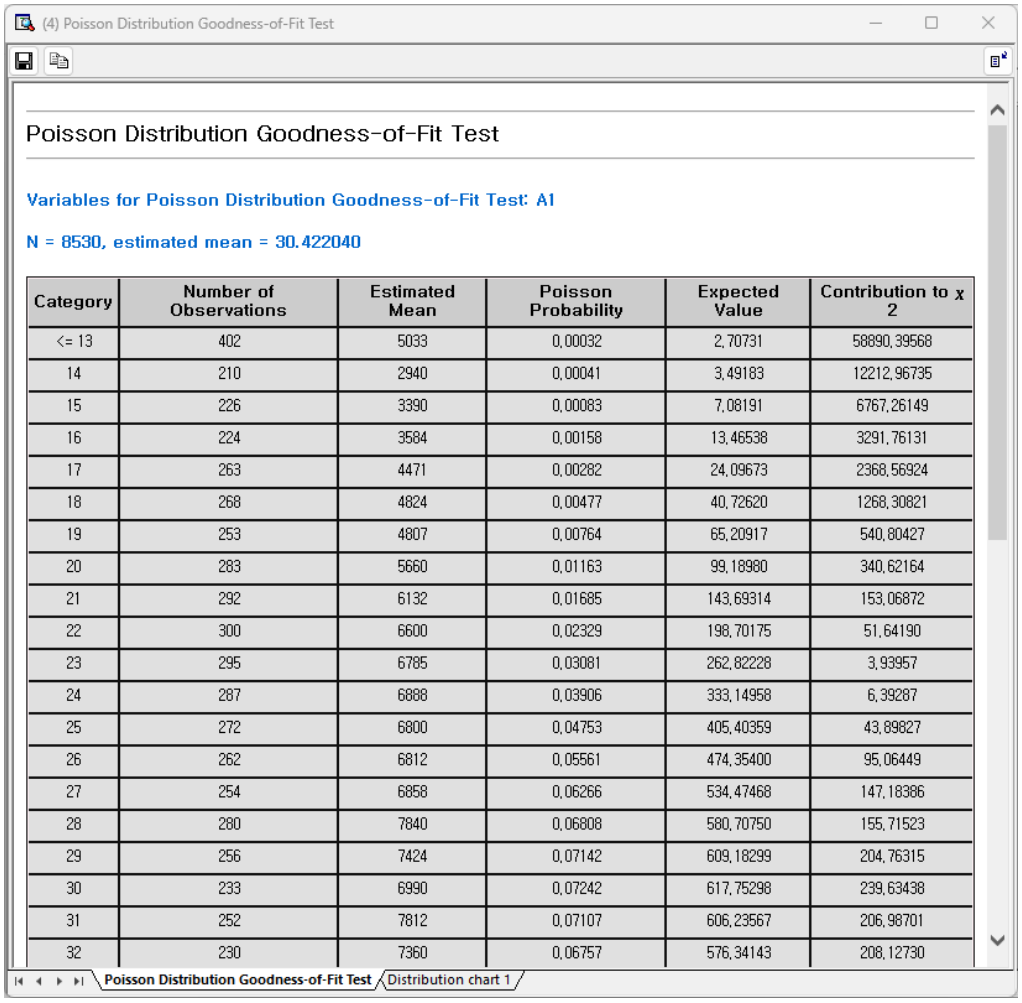
[Analyze] – [Basic Statistics] – [Poisson Test]

Select a variable. If you want to a distribution chart of observed and expected values, then select it.



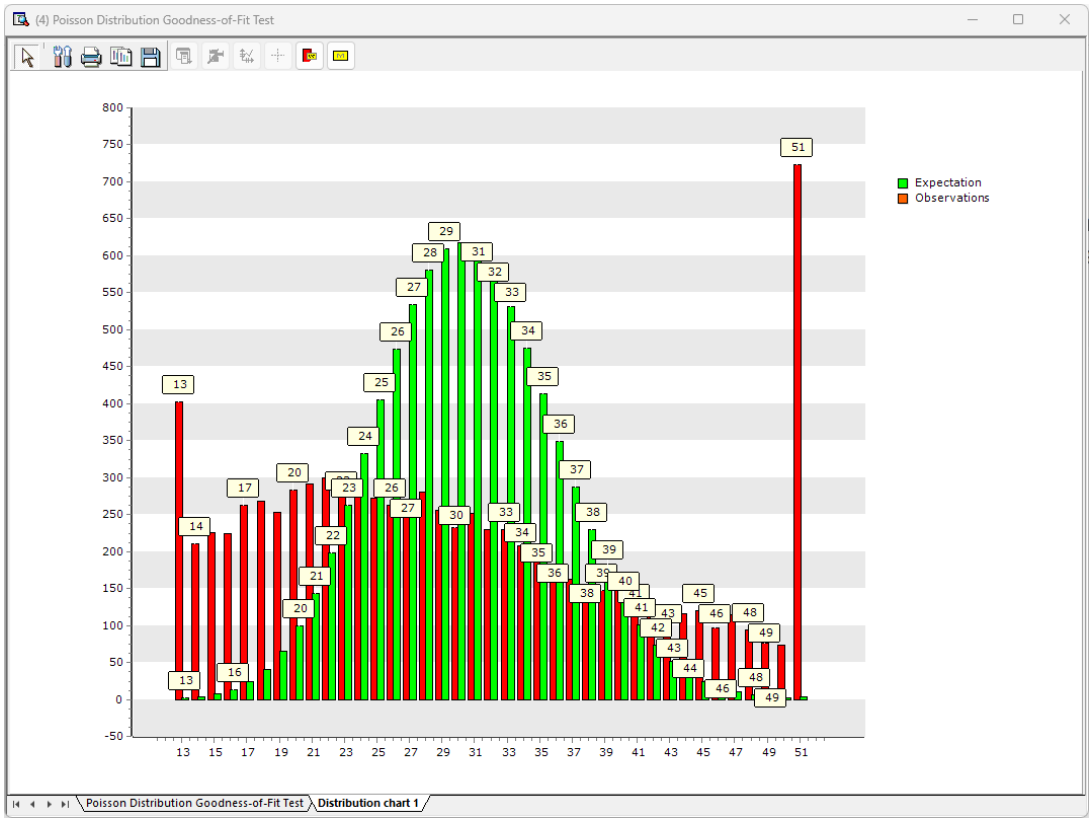
Results

The analysis results include the Poisson probability, an expected value, and chi-square test statistic for each variable.



Category	Number of Observations	Estimated Mean	Poisson Probability	Expected Value	Contribution to χ^2
<= 13	402	5033	0,00032	2,70731	58890,39568
14	210	2940	0,00041	3,49183	12212,96735
15	226	3390	0,00083	7,08191	6767,26149
16	224	3584	0,00158	13,46538	3291,76131
17	263	4471	0,00282	24,09673	2368,56924
18	268	4824	0,00477	40,72620	1268,30821
19	253	4807	0,00764	65,20917	540,80427
20	283	5660	0,01163	99,18980	340,62164
21	292	6132	0,01685	143,69314	153,06872
22	300	6600	0,02329	198,70175	51,64190
23	295	6785	0,03081	262,82228	3,93957
24	287	6888	0,03906	333,14958	6,39287
25	272	6800	0,04753	405,40359	43,89827
26	262	6812	0,05561	474,35400	95,06449
27	254	6858	0,06266	534,47468	147,18386
28	280	7840	0,06808	580,70750	155,71523
29	256	7424	0,07142	609,18299	204,76315
30	233	6990	0,07242	617,75298	239,63438
31	252	7812	0,07107	606,23567	206,98701
32	230	7360	0,06757	576,34143	208,12730

In the Poisson distribution tab of the results window, you can view the distribution chart of observed and expected values.



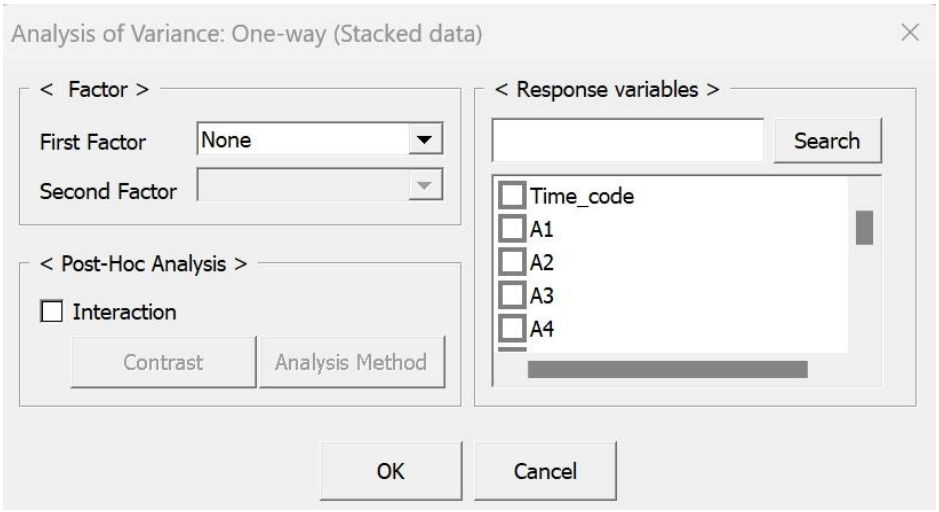
4.3.3 Variance Analysis

4.3.3.1 One-way ANOVA

One-way ANOVA is a statistical method that involves one factor and is used to determine whether there are significant differences in the means of three or more independent groups.

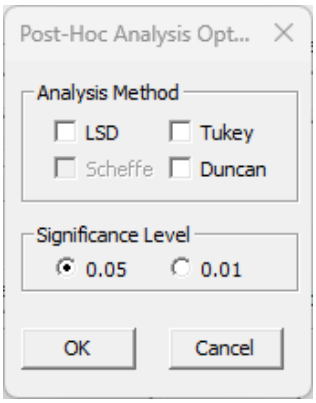
How to run

[Analyze] - [Variance Analysis] - [One-way ANOVA]



Factor: Select one discrete variable which has different groups or categories.

Observed Value: Select continuous variables that affect the group means.



Post-Hoc analysis: Select the analysis method which identify which groups differ from each other. Moreover, select desired significance level depending on your needs.

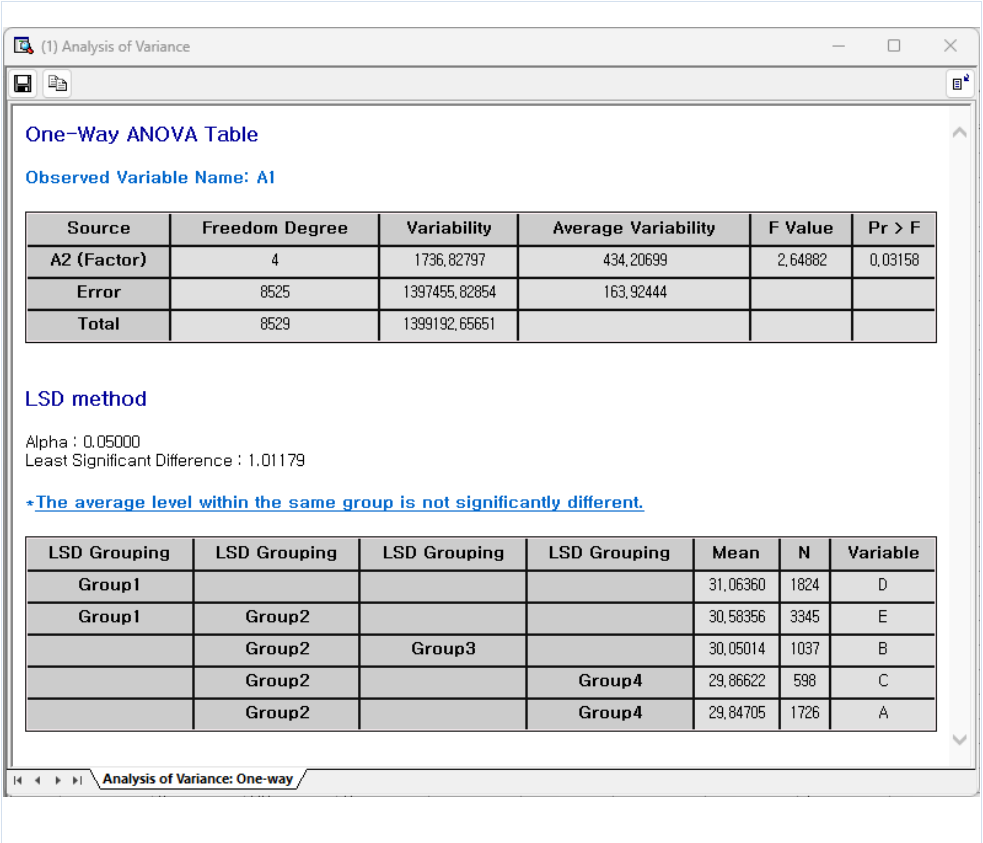
LSD: more sensitive to small differences than **Tukey**

Tukey: robust and reliable than **LSD**

Duncan: useful when larger pairs of means are being compared.

Result

Results can be saved to an HTML file by clicking the save icon located in the top-left corner of the window.

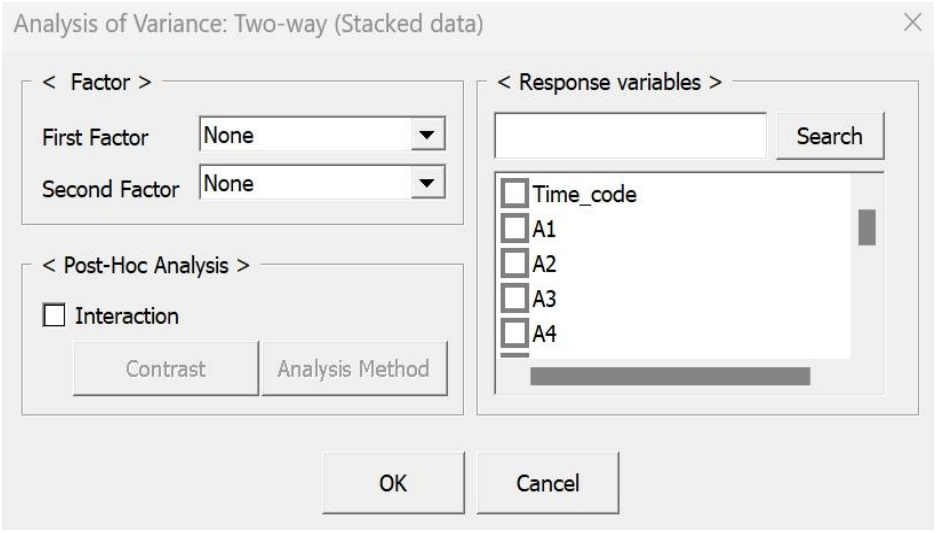


4.3.3.2 Two-way ANOVA

Two-way ANOVA is a statistical method that involves two factors and interaction between the two factors. It is used to determine whether there are significant differences in the means of three or more independent groups.

How to run

[Analyze] - [Variance Analysis] - [Two-way ANOVA]



First/Second Factor: Select two discrete variables that which has different groups or categories.

Observed Value: Select continuous variables that affect the group means.

Result

Results can be saved to an HTML file by clicking the save icon located in the top-left corner of the window.

Two-Way ANOVA Table

Observed Variable Name: result

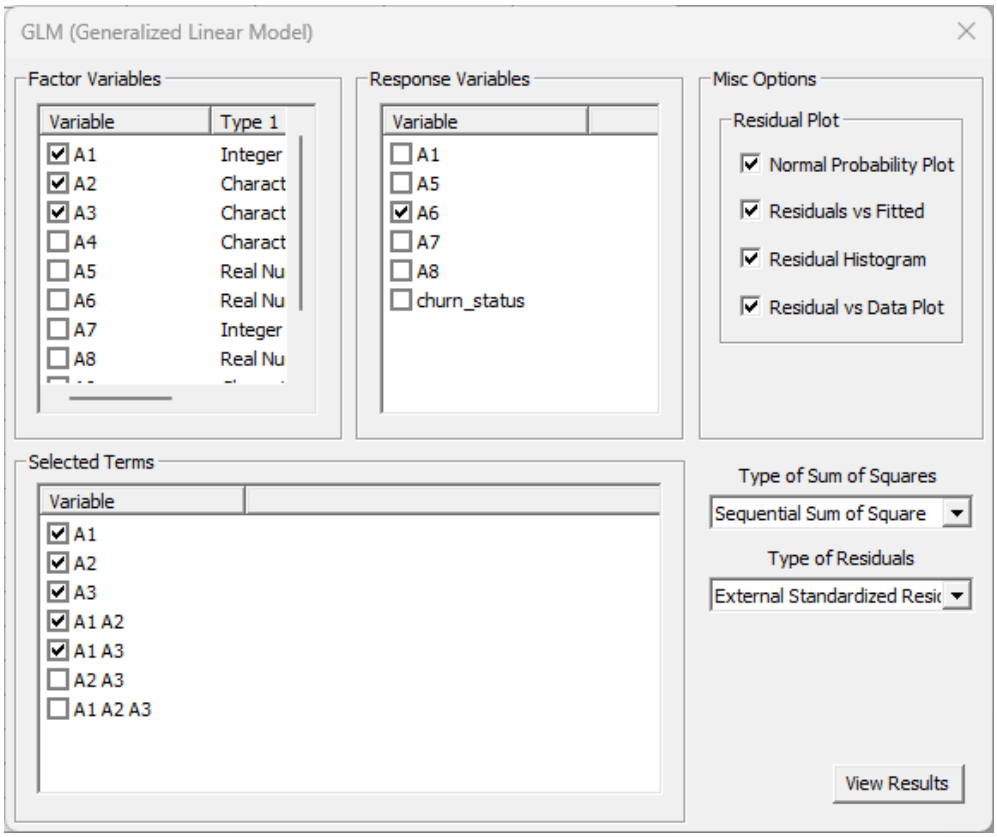
Source	Freedom Degree	Variability	Average Variability	F	P value
brightness (Factor A)	2	100,00000	50,00000	0,30233	0,74416
temperature (Factor B)	2	5200,00000	2600,00000	15,72093	0,00034
Residual Error	13	2150,00000	165,38462		
All	17	7450,00000			

4.3.3.3 GLM (General Linear Model)

General Linear Model can handle imbalanced data that often appears in real situations and can also handle covariates.

How to run

[Analyze] – [Variance analysis] – [General Linear Model]



Factor Variables: Enter the variables corresponding to the factors used in the experiment. At this time, the selected continuous variable is automatically processed as a covariate.

Response Variables: Multiple selections are possible, and if multiple selections are made, you can obtain results for analysis of variance and linear regression models for multiple response variables.

Selected Terms: Select interaction terms you may want.

Residual Plot: Selected residual plots will be displayed on the results page.

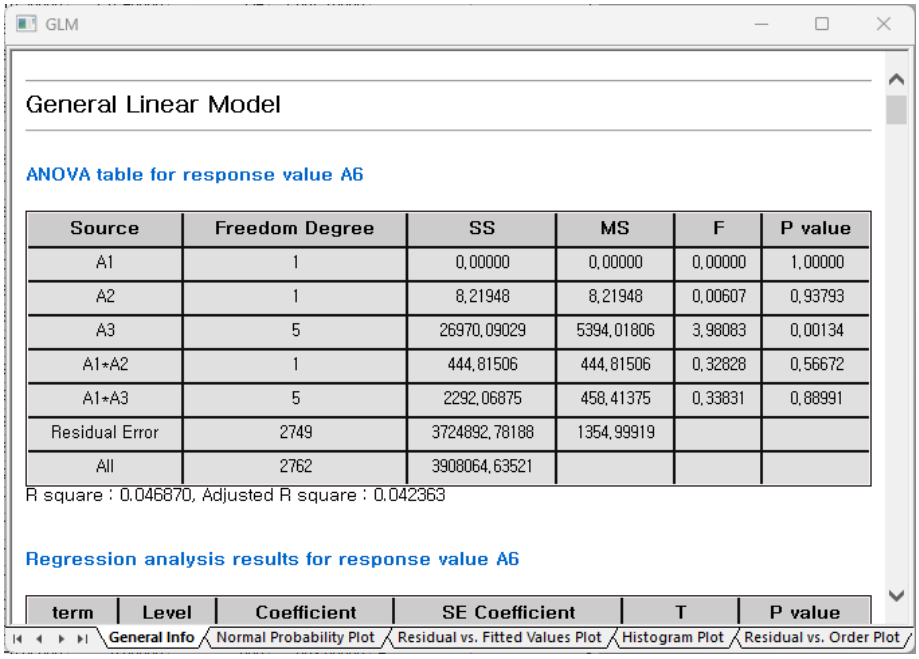
Type of sum of Squares: Choose whether to use **Modified** or **Sequential Sum of Square**.

Type of Residuals: Choose whether to use **Residuals**, **Standardized Residuals**, or **External Standardized Residuals**.

Result

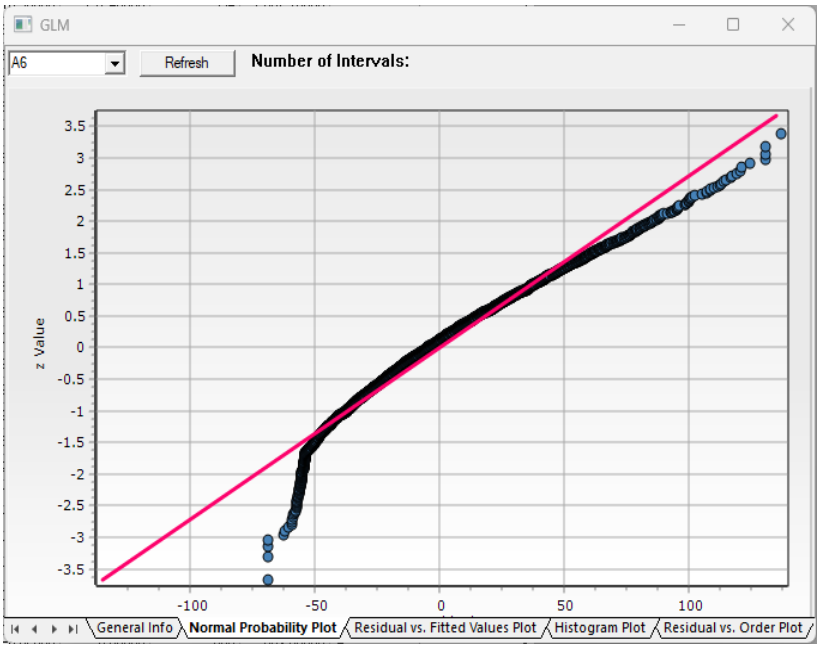
General Information shows the ANOVA Table, regression analysis results, and abnormal

observation.



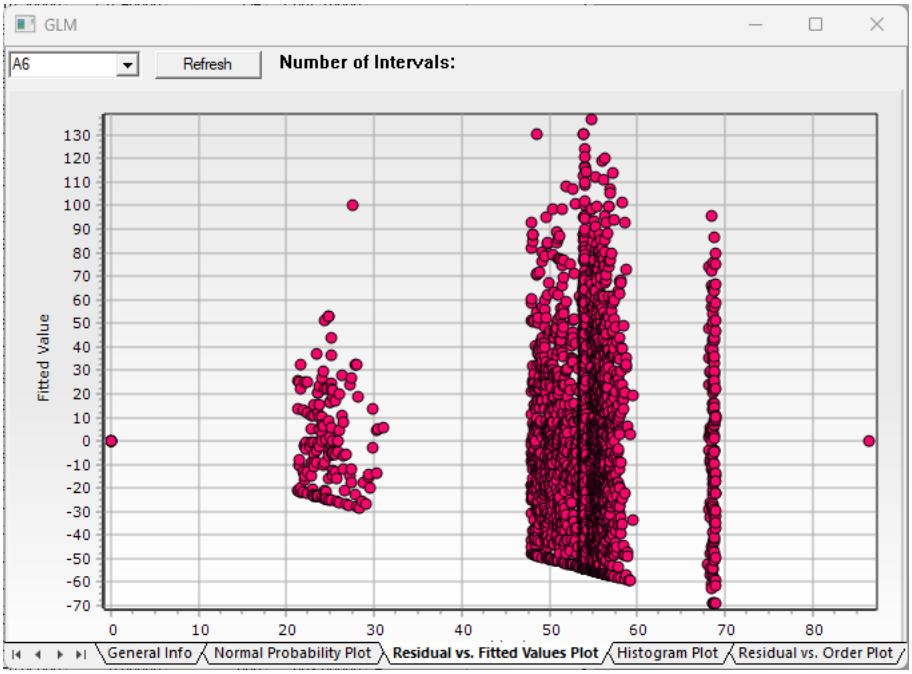
- **Normal Probability Plot**

  More data points near the red line indicate that the residuals are closer to a normal distribution.



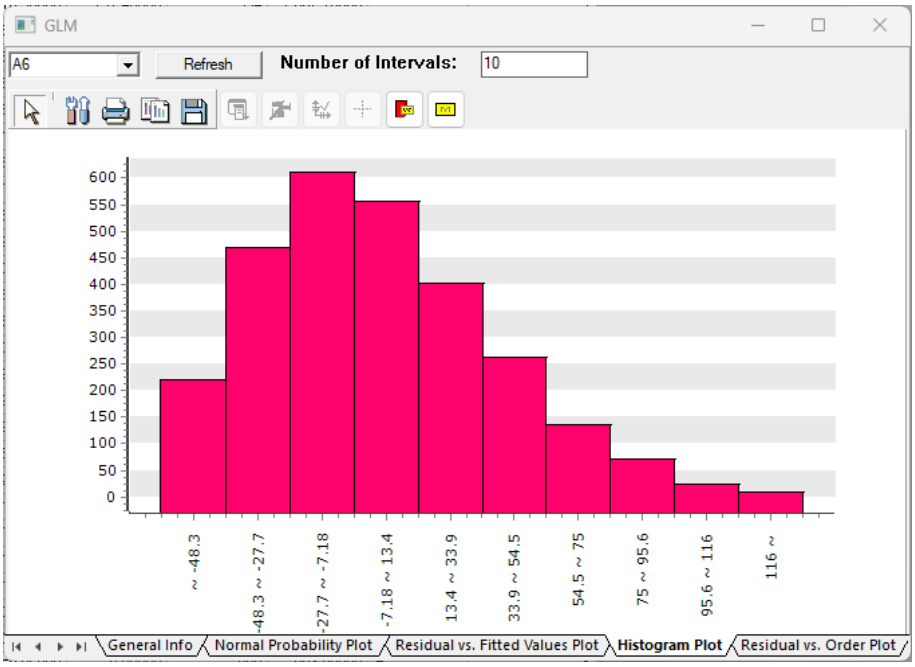
- **Residual vs. Fitted Values Plot**

Plot with residuals on the horizontal axis and fitted values on the vertical axis.



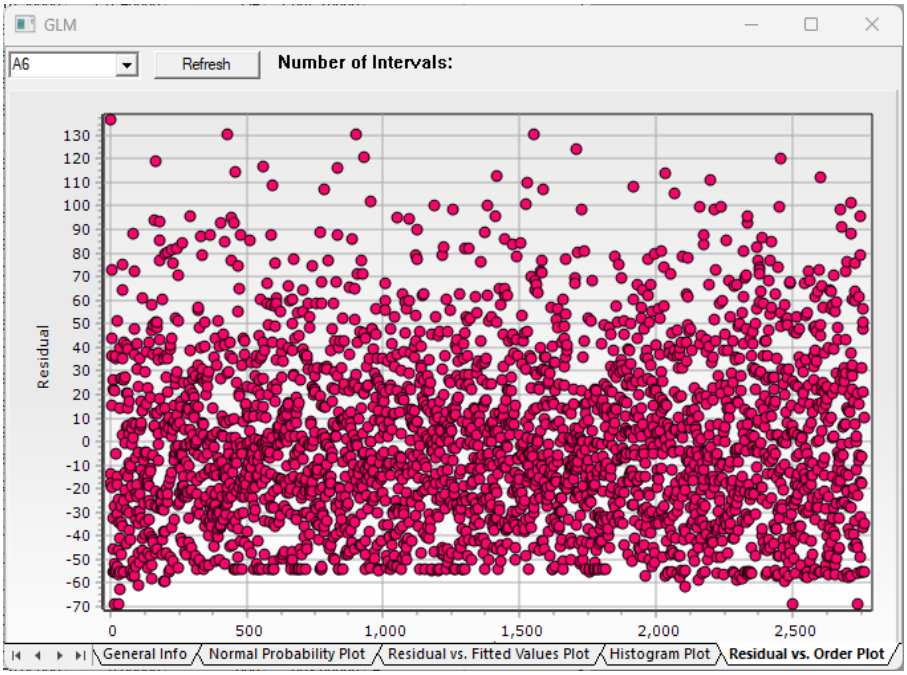
- **Histogram Plot**

  Distribution of residuals



- **Residual vs. Order Plot**

  Residuals to the order of the data points.



4.3.4 Regression Analysis

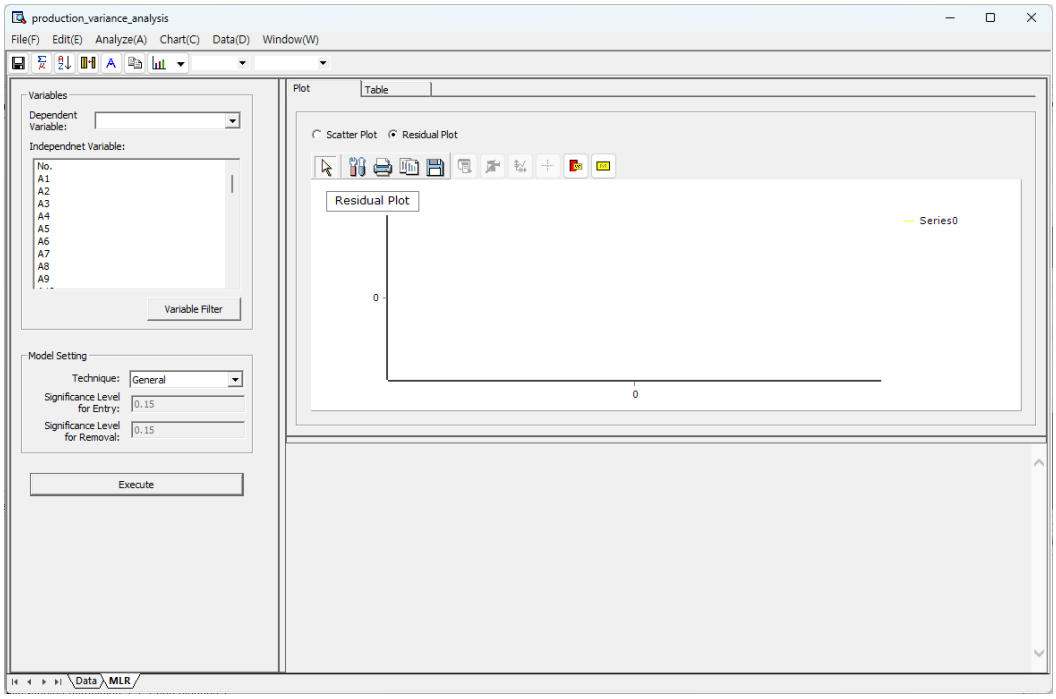
4.3.4.1 Multiple Linear Regression

Multiple Linear Regression predicts the dependent variable through the linear equation of two or more independent variables. MLR settings can be chosen between Stepwise or General where Stepwise technique combines Forward, Backward methods to iteratively add or remove variables, refining the model with significant predictors.

How to run

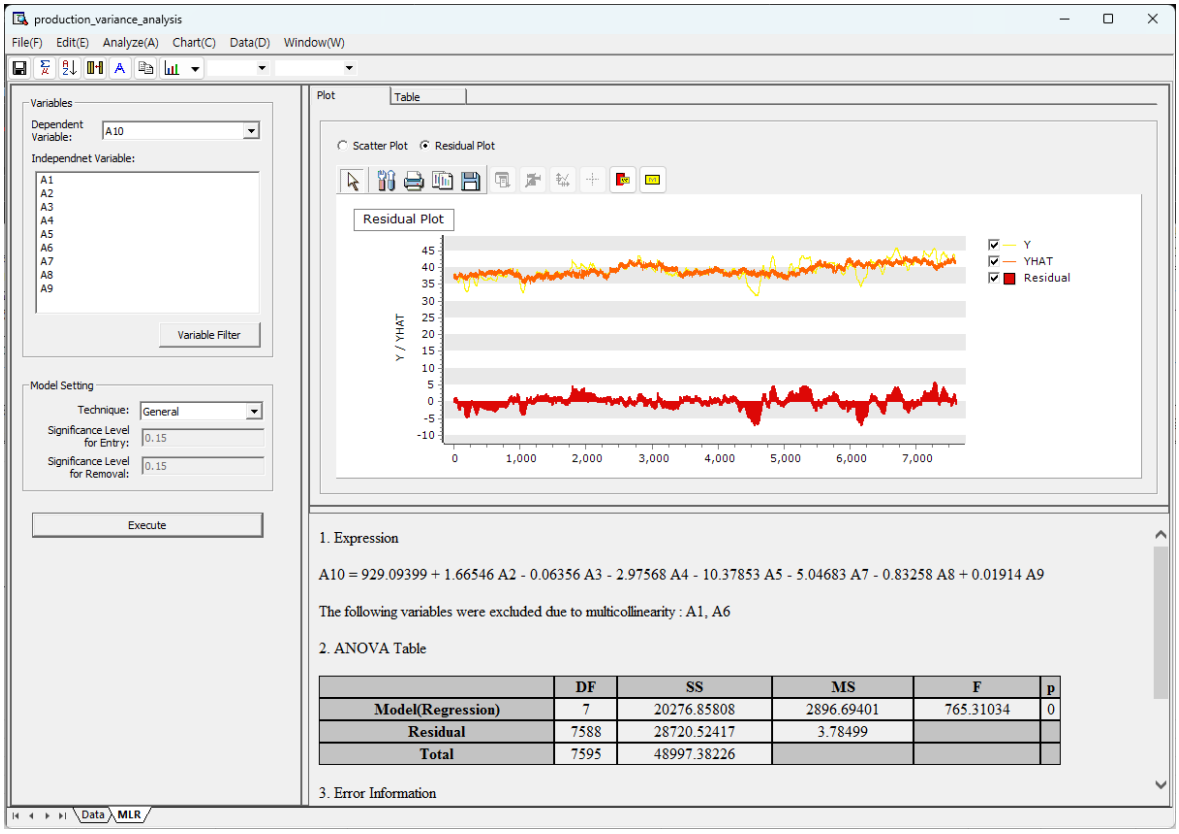
[Analyze] – [Regression Analysis] – [Multiple Linear Regression]

Select Dependent Variable from the dropdown selection box and Independent Variables through clicking the Variable Filter button. Then in the Model Setting, choose the Multiple Linear Regression technique between General and Stepwise. When Stepwise is chosen, also set the options of significance levels. Then click Execute button to see the results.

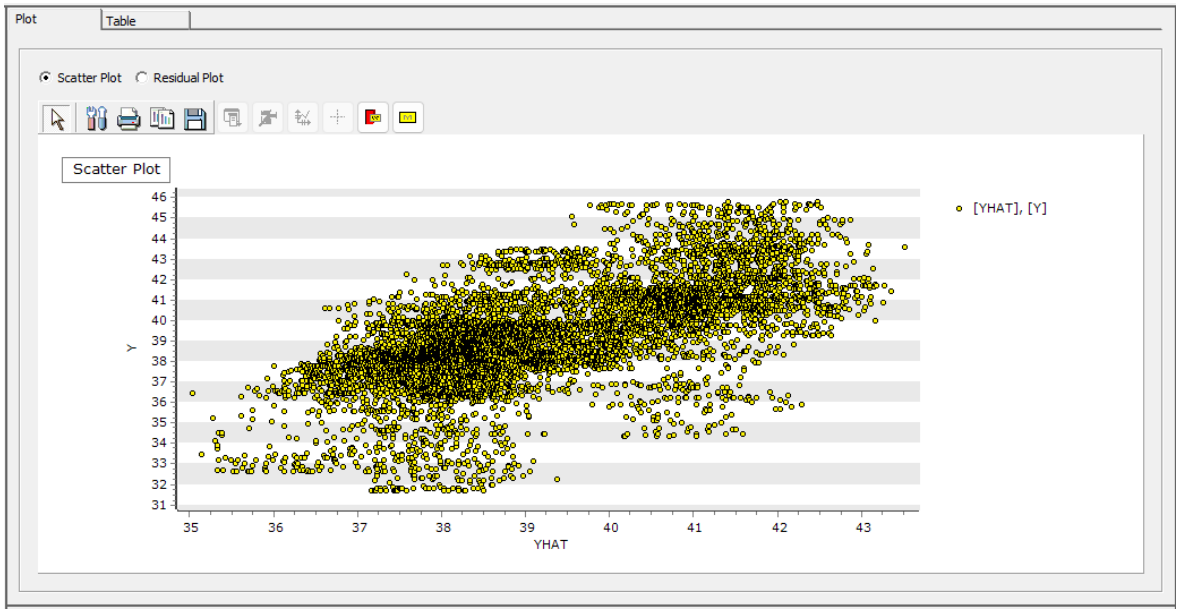


Results

The analysis results include Residual/Scatter Plots and MLR result table, where each result can be viewed through Plot and Table tab, respectively.



	DF	SS	MS	F	p
Model(Regression)	7	20276.85808	2896.69401	765.31034	0
Residual	7588	28720.52417	3.78499		
Total	7595	48997.38226			



Plot Table

	2	3	4	5	6	7	8	9	10	11	12	13
	A2	A3	A4	A5	A6	A7	A8	A9	A10	MLR_YHAT	MLR_RES	MLR_LEV
1	1,95173	115,15100	27,27380	72,20820	109,96700	0,55476	66,99480	75,06930	38,00220	37,31089	0,69131	0,00192
2	1,97996	115,48600	27,05320	72,20820	109,96700	0,55524	66,99330	75,05090	37,96510	37,99151	-0,02641	0,00181
3	2,02079	115,31600	27,05140	72,20820	110,07900	0,55242	67,05090	75,06930	37,98350	38,04229	-0,05879	0,00184
4	2,00120	115,33800	27,31790	72,20820	109,96700	0,55388	66,93840	75,05210	37,98350	37,30121	0,68229	0,00148
5	2,02360	115,19200	27,04110	72,20820	109,96700	0,55435	66,99420	75,14410	37,98350	38,12444	-0,14094	0,00191
6	2,02063	115,62000	27,29520	72,20820	109,96700	0,55059	67,08870	75,05210	37,98350	37,27469	0,70881	0,00149
7	1,99856	115,26400	27,22800	72,22000	109,96700	0,55992	66,99470	75,05090	37,98350	37,36920	0,61430	0,00174
8	2,02636	115,48900	27,15290	72,27770	110,07900	0,55333	66,99420	74,95700	37,98350	37,05772	0,92578	0,00342
9	2,01522	115,46700	27,35910	72,25440	109,96700	0,55808	66,96790	74,95670	38,00220	36,66672	1,33548	0,00249
10	2,00764	115,39400	27,08530	72,23120	109,94900	0,55435	67,09010	75,14410	37,98380	37,63495	0,34885	0,00217
11	1,98164	115,31500	27,11020	72,23140	109,96700	0,55759	66,86390	75,04900	38,05830	37,69065	0,36765	0,00217
12	1,99189	114,94700	27,02340	72,23140	110,06000	0,55384	67,02280	74,93680	38,05830	37,87387	0,18443	0,00235
13	1,97797	114,98400	27,05420	72,23120	109,96700	0,55715	66,94860	74,71240	38,17050	37,79953	0,37097	0,00229
14	1,97222	115,30400	27,34290	72,21980	110,07900	0,55242	67,04990	74,20900	38,17050	36,95877	1,21173	0,00164
15	1,95748	115,33600	27,14210	72,21980	109,96700	0,55855	66,94800	73,79750	38,17050	37,57573	0,59477	0,00189
16	2,13262	115,15300	27,13500	72,23120	110,07900	0,55711	66,99650	73,79750	38,05830	37,74873	0,30957	0,00503
17	1,96840	116,10400	27,04340	72,21980	109,96700	0,55341	66,91980	73,49830	38,05830	37,88249	0,17581	0,00186
18	1,94347	114,64800	27,22280	72,21980	109,96700	0,55710	67,05040	73,38610	38,11530	37,27015	0,84515	0,00216
19	1,96301	116,02800	27,02510	72,23140	109,96700	0,55246	66,93880	73,29230	38,05830	37,79746	0,26084	0,00216
20	1,99941	115,02400	27,22800	72,23140	109,96800	0,55663	67,02320	73,23710	38,05830	37,22571	0,83259	0,00167
21	1,99473	115,55900	27,12200	72,21980	109,96700	0,55290	66,99370	72,97500	38,07670	37,65811	0,41859	0,00143

Below, the regression equation, ANOVA table, and Standard Error Information table are displayed.

1. Expression

A10 = 929.09399 + 1.66546 A2 - 0.06356 A3 - 2.97568 A4 - 10.37853 A5 - 5.04683 A7 - 0.83258 A8 + 0.01914 A9

The following variables were excluded due to multicollinearity : A1, A6

2. ANOVA Table

	DF	SS	MS	F	p
Model(Regression)	7	20276.85808	2896.69401	765.31034	0
Residual	7588	28720.52417	3.78499		
Total	7595	48997.38226			

3. Error Information

R-squared	0.41384
Adjusted R-squared	0.41329
RMSE	1.94448
MAE	1.44856
MAPE	3.74351

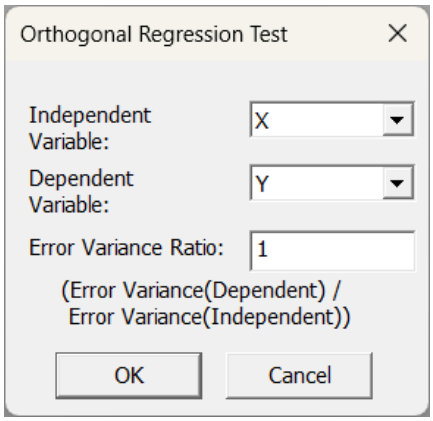
4.3.4.2 Orthogonal Regression Analysis

Orthogonal regression assumes errors in both independent and dependent variables and minimizes the sum of squared orthogonal distances between the observations and the regression line. Unlike ordinary least squares (OLS), which only considers errors in the dependent variable, orthogonal regression accounts for errors in both variables, yielding a more accurate regression line.

How to run

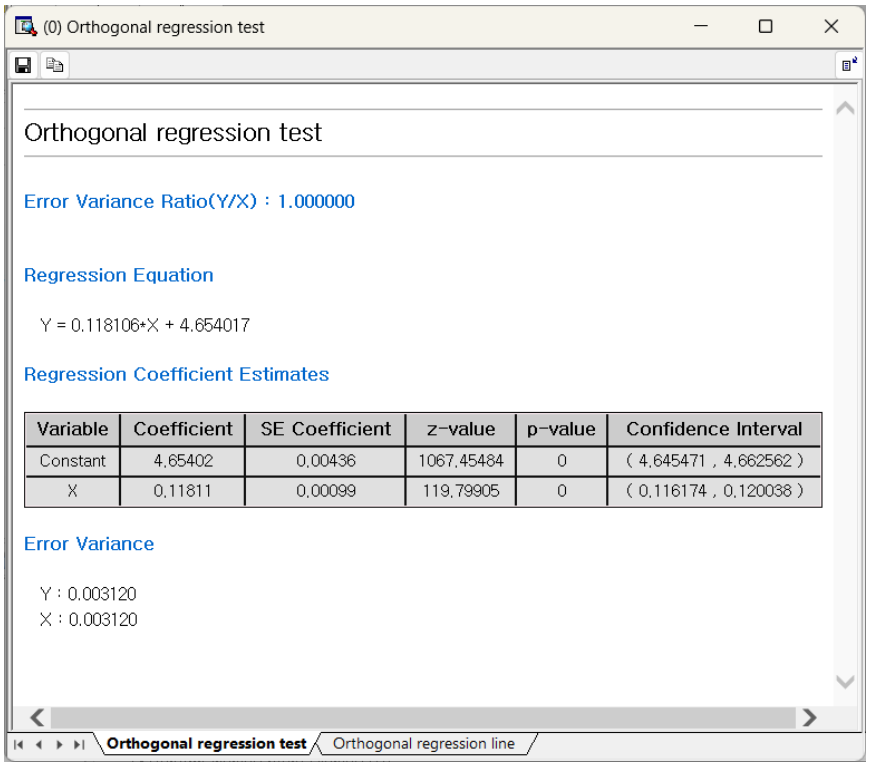
[Analyze] – [Regression Analysis] – [Orthogonal Regression Analysis]

Select Independent Variable and Dependent Variable and input Error Variance Ratio (Variance of the Dependent Variable's Error/ Variance of the Independent Variable's Error).

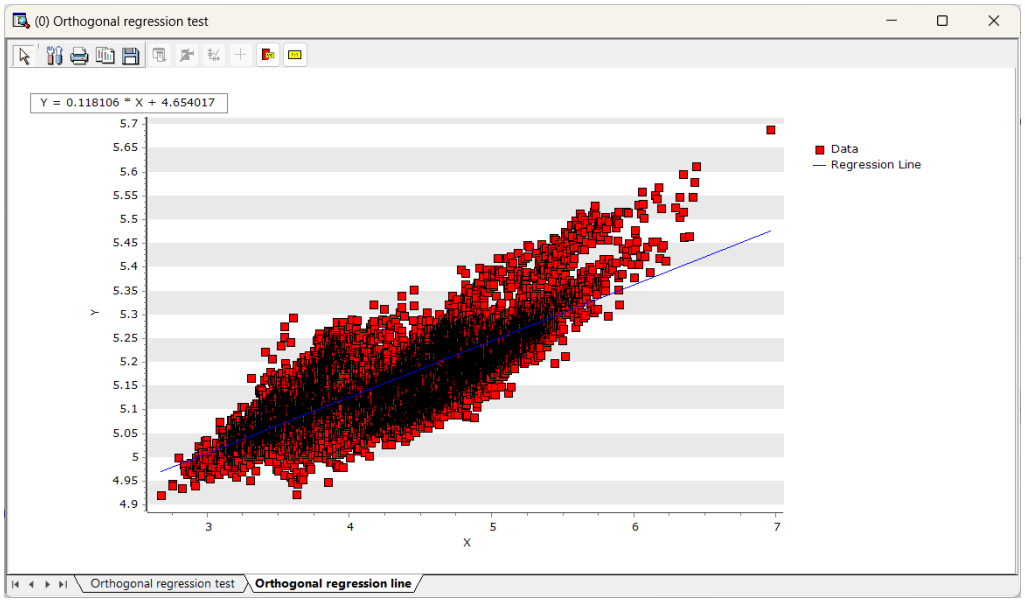


Results

The analysis results include the orthogonal regression equation, a regression coefficient estimates, and the error variance for each variable.



In the Orthogonal regression line tab of the results window, you can view the graph of the observed values and the orthogonal regression line.



4.3.4.3 Nonlinear Regression

Overview

Nonlinear Regression is applied when the relationship between the independent and dependent variables is **nonlinear**, with the regression equation expressed as a nonlinear function (such as polynomial, exponential, or logarithmic functions).

- **Nonlinear Least Square Regression**

$$F(\mathbf{x}) = \frac{1}{2}\sum_{i=1}^{n}(y_i - M(\mathbf{x}|t_i))^2 = \frac{1}{2}\sum_{i=1}^{n}(f_i(\mathbf{x}))^2 = \frac{1}{2}\| \mathbf{f}(\mathbf{x}) \|^2$$

ECMiner™ Nonlinear Regression uses **the Levenberg-Marquardt algorithm**, a high-performance method widely used for nonlinear regression and optimization problems. The Levenberg-Marquardt algorithm performs an optimization process to minimize the **Nonlinear Least Square Regression** (as shown in the equation above) and efficiently adjusts parameters by combining the Gauss-Newton method and gradient descent.

- **Curve Fitting in ECMiner™**

  Standard Expression

: ECMiner™ supports various standard regression analyses that can be linearized for parameter estimation and generally provides accurate predictions.

- Polynomial Function
- Logarithm Function
- Exponential Function
- Power Law Function
- Rational Function

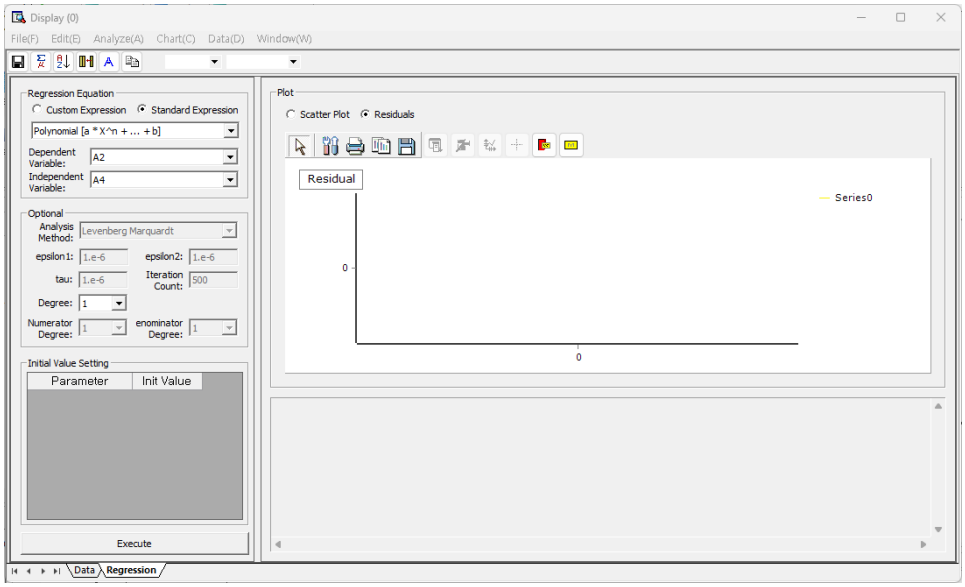
Custom Expression

: If you input a desired function shape, the algorithm will find the best-fitting parameters.

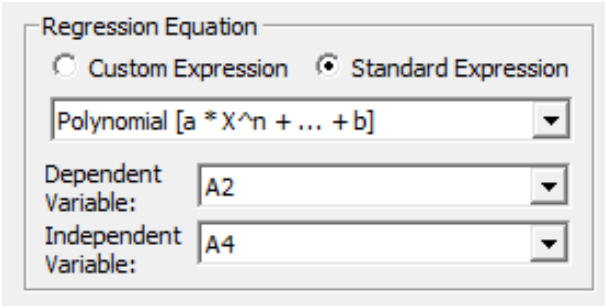
How to run

[Analyze] – [Regression Analysis] – [Nonlinear Regression Analysis].

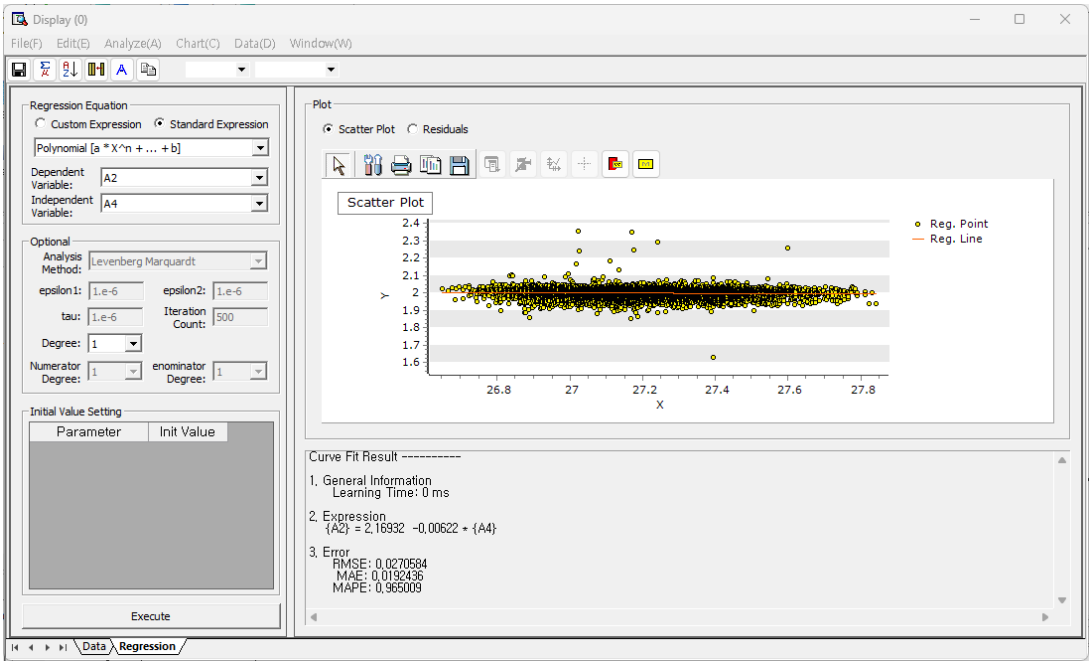
Specify either Custom Expression or Standard Expression.



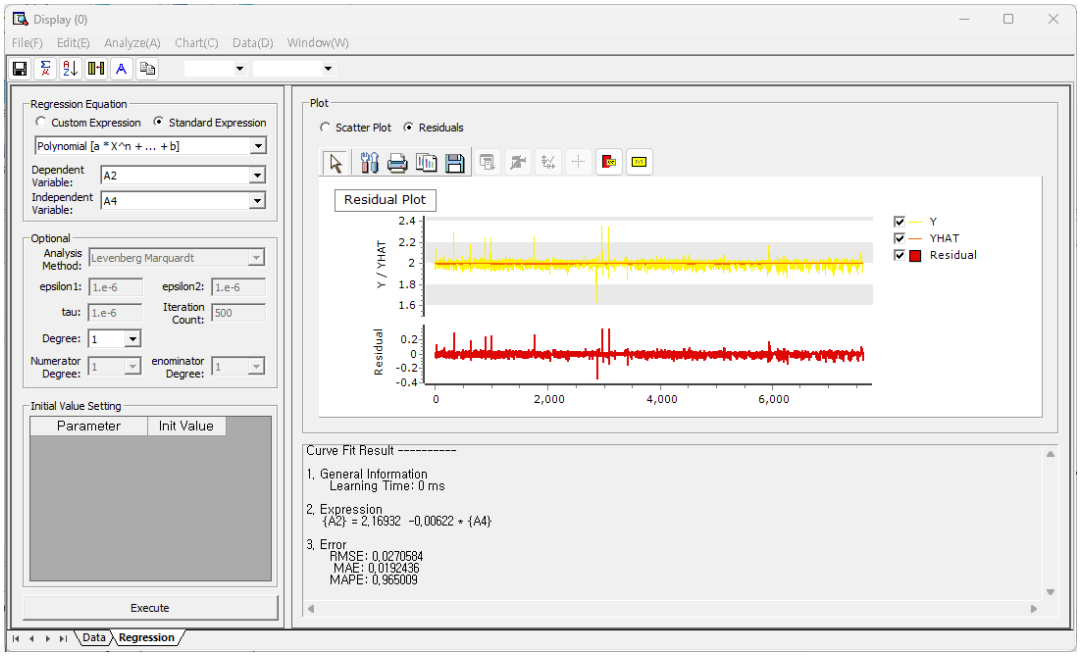
If you select the Standard Expression, define the independent and dependent variables.



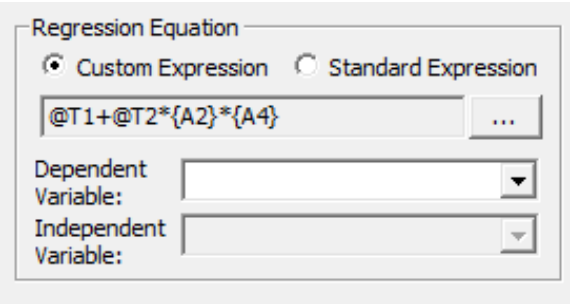
When you specify variables, select the form of the function, and press the Run button, you can view the estimated results (plot and regression analysis results) as follows.



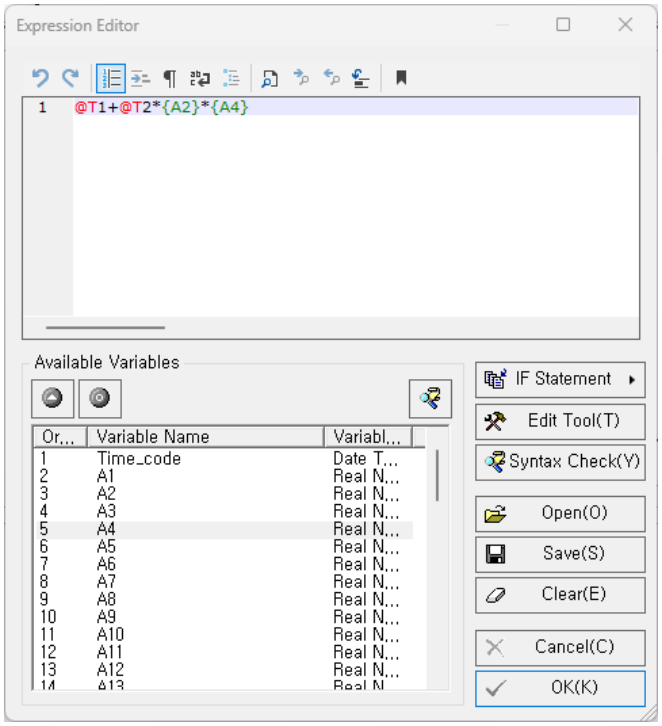
If you select Residuals, the scatterplot of the time series data and fitted values. You can select different types of residuals.



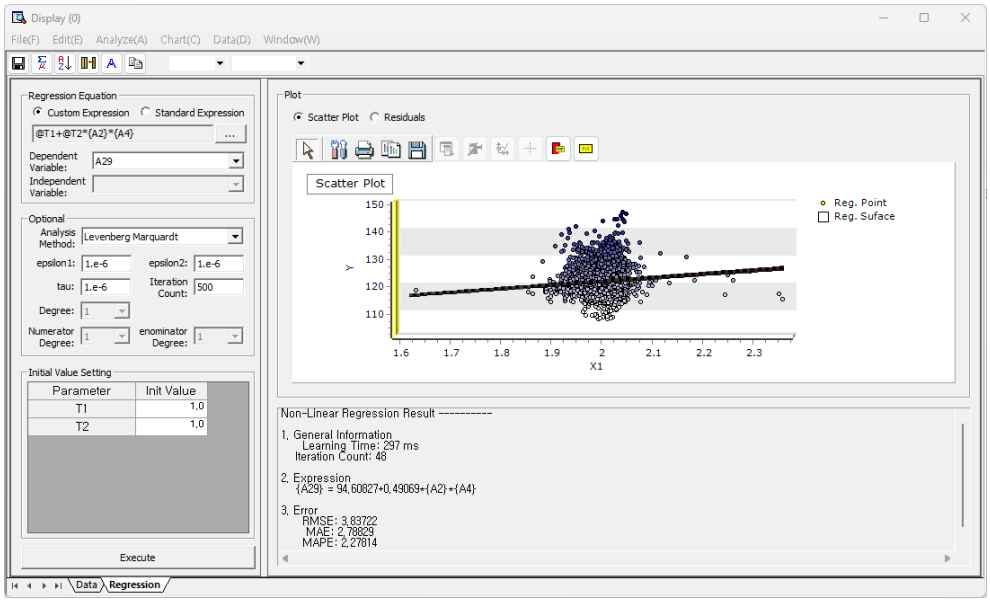
If you select the Custom Expression, click the ... button and specify the formula you want.



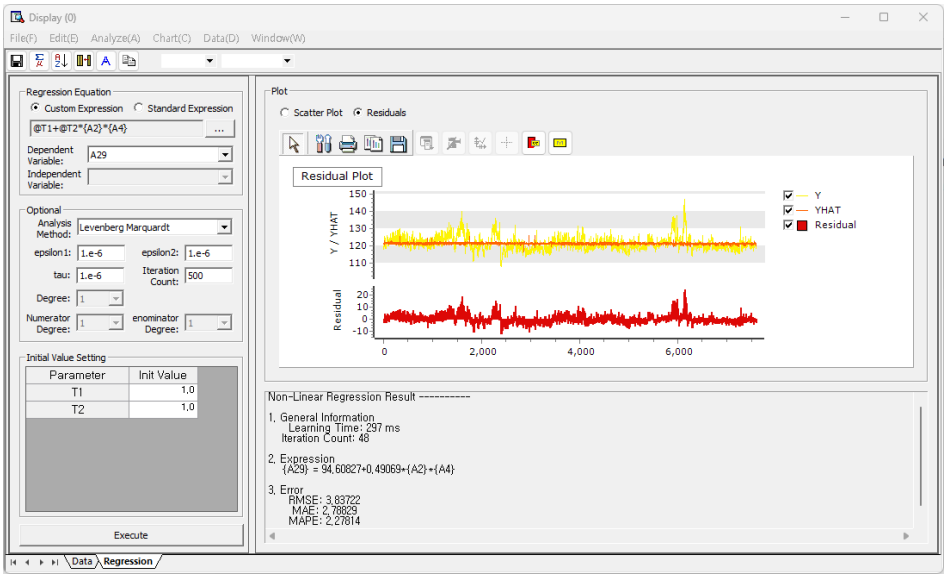
(At this time, please note that the name of the parameter to be estimated must be written in the form @T1, @T2, etc.)



If there are two variables used as independent variables above, the following 3D plot can be obtained as a result.



If you select the Residuals radio button, you can view the actual time series data, fitted values, and residual trends as follows.



4.3.5 SPC (Statistical Process Control)

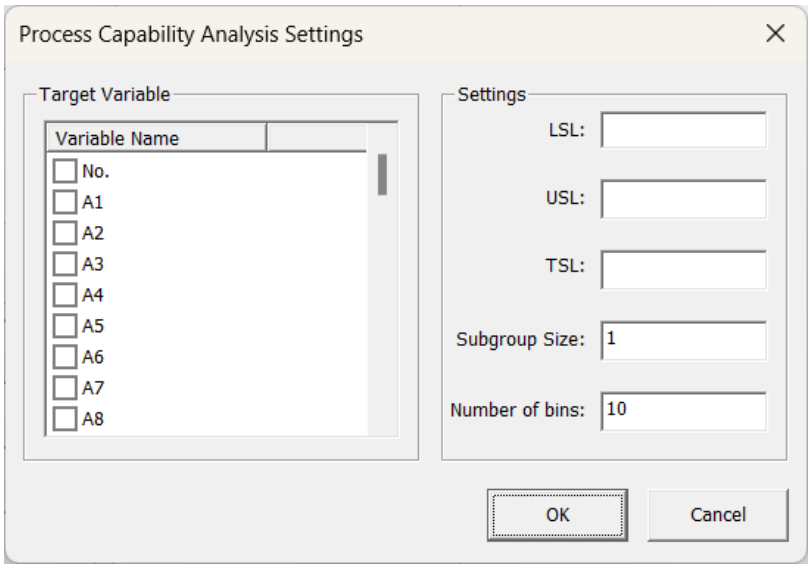
4.3.5.1 Process Capability Analysis

Process Capability Analysis is for evaluating whether data is distributed within the desired specification.

How to run

[Analyze] - [SPC] - [Process Capability Analysis]

Click [Settings] button, and select the target variable.



Select Target Variable (multiple selections are possible) and set the LSL, USL, TSL, Subgroup Size, and Number of bins or histograms.

Results

Terminology	Description	etc.
Histogram	Histogram	
Intracluster	Normal distribution curve using the mean of and the standard deviation within the subgroup.	The chart varies by **group size**
All	Normal distribution curve using the mean and standard deviation	
LSL	Lower specification limit. The lower limit on the process specification.	It depends on **LSL** in the **[Settings]** window.
USL	Upper specification limit. The upper limit on the process specification	It depends on **USL** in the **[Settings]** window.
LCL	Lower control limit. LCL is used as a threshold (LCL = mean - 3 * standard deviation)	
UCL	Upper control limit. UCL is used as a threshold (UCL = mean + 3 * standard deviation)	
Cp	Process Capability Index (PCI). PCI is a criterion used to evaluate process capability. Cp = (USL - LSL)/(6 * within-group standard deviation)	
CpL	CpL is a measure of how close the subgroup average is to the LSL. The lower the CpL, the higher the likelihood of producing defects. CpL = (mean – LSL)/(3 * within-group standard deviation)	
CpU	CpU is a measure of how close the subgroup average is to the USL. The lower the CpU, the higher the likelihood of producing defects. CpU = (USL – mean)/(3 * within-group standard deviation)	
Cpk	Cpk considers the shift in the process mean. Unlike Cp, Cpk measures how well a process is centered between specification limits	
Pp	Process capability index (PCI) for the entire process without considering subgroups. Pp = (USL – LSL)/(6 * standard deviation)	
PpL	Overall average of the process is to the LSL.	

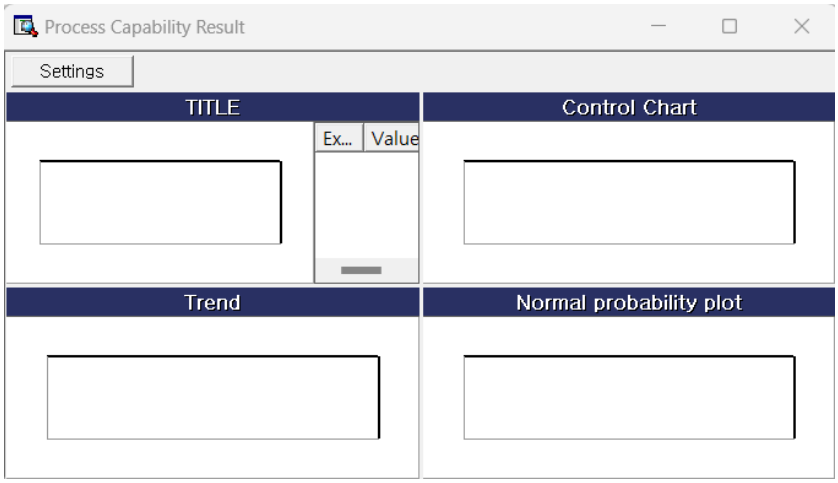
	PpL = (mean – LSL)/(3 * standard deviation)	
PpU	PpU is a measure of how close the overall average of the process to the USL. PpU = (USL – mean)/(3 * standard deviation)	
Ppk	Ppk considers the shift in the overall process mean. Unlike Pp, Ppk measures how well a process is centered between specification limits	
Cpm	A Process Capability Index (PCI) that indicates how close the process is to the target value. Calculated as the deviation from the process average and target value	

4.3.5.2 Process Capability Result

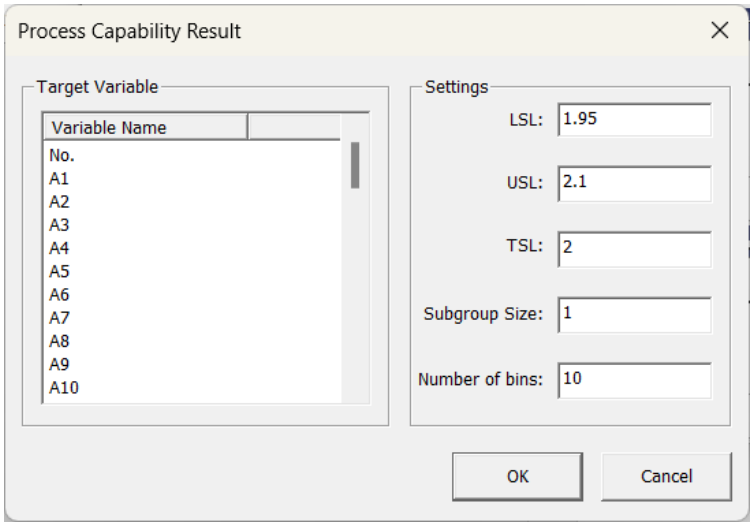
Process Capability Summary provides summary report for a specific field whether data is distributed within the desired process specification area.

How to run

[Analyze] - [SPC] - [Process Capability Result]

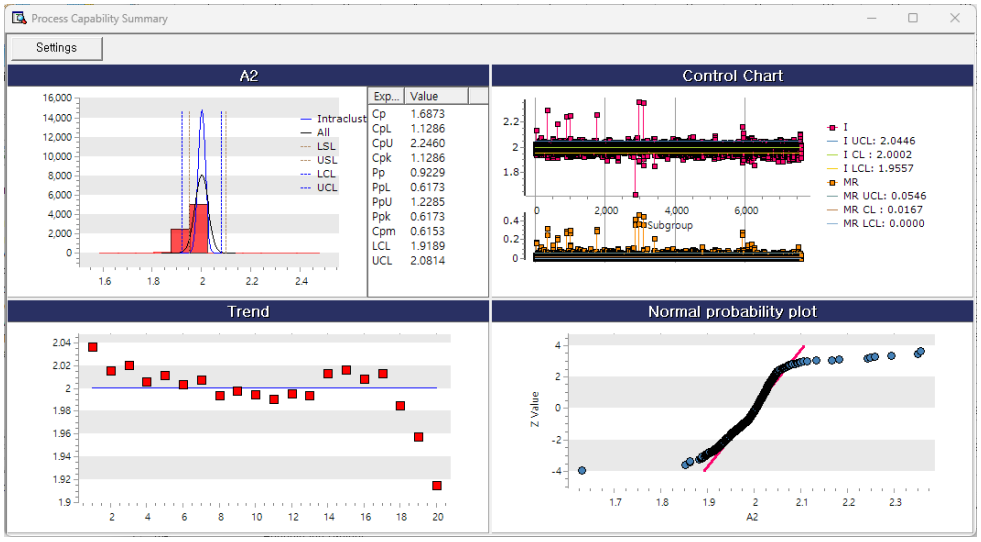


Click [Settings] button, and select the target variable.



Select Target Variable and set the LSL, USL, TSL, Subgroup Size, and Numbers of bins for multiple charts.

Results



4.3.5.3 Acceptance Sampling

Acceptance Sampling is a technique that determines whether a lot accept or fails a sample product randomly drawn from the lot. If you are counting defective or non-defective items, choose Attributes Acceptance Sampling. If you are testing a variable against a specific threshold to

determine whether it is defective, choose Variable Acceptance Sampling.

---

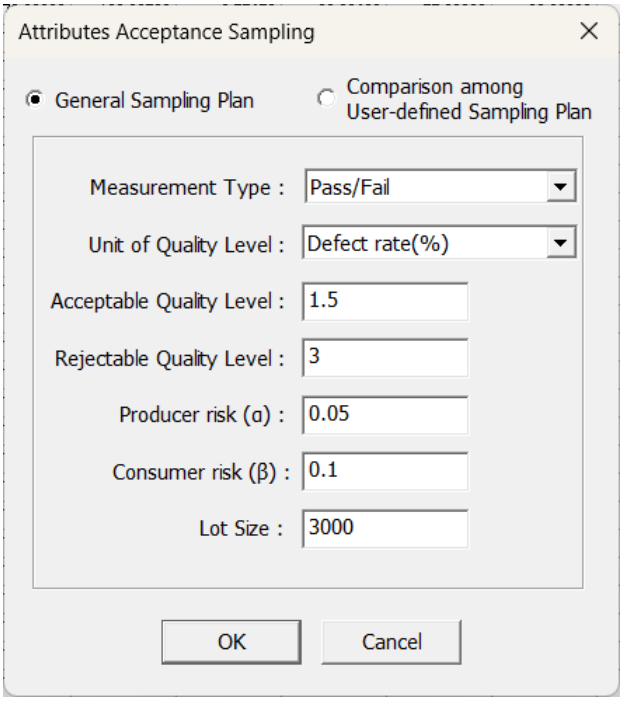
(1) Attributes Acceptance Sampling

How to run

[Analyze] – [SPC] – [Acceptance Sampling] – [Attributes Acceptance Sampling],

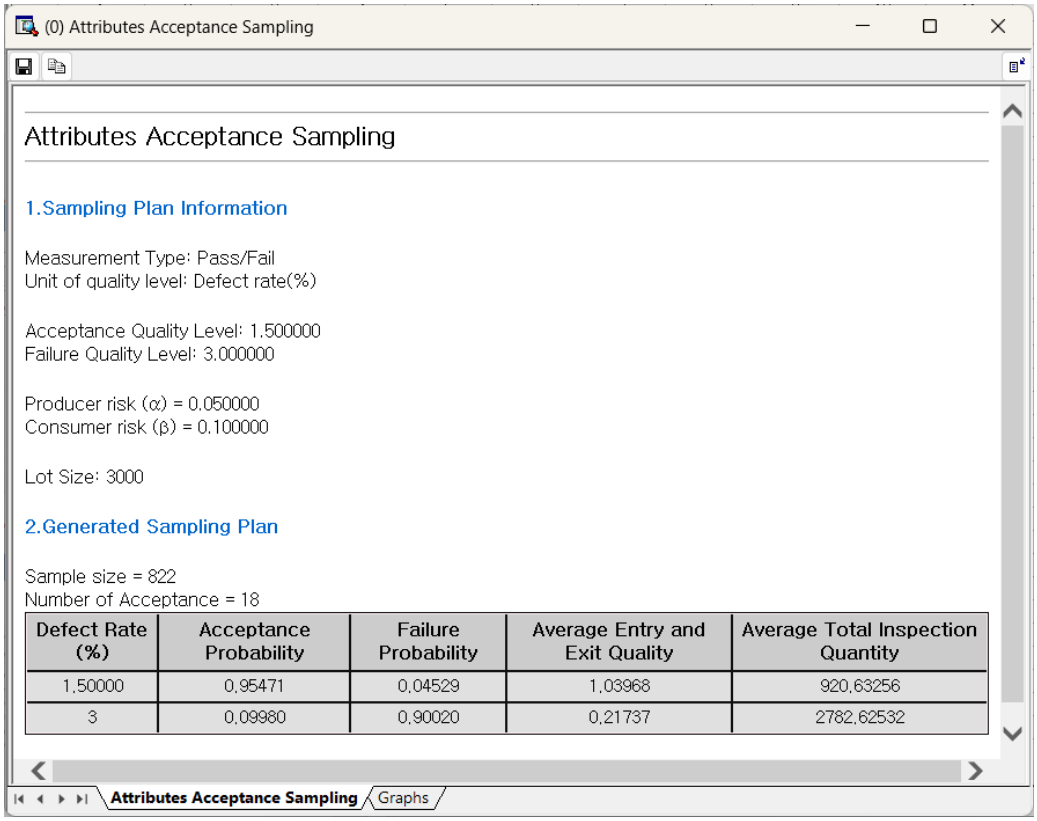
Select General Sampling Plan. The result gives you the optimal sample size and number of acceptances. If you want to compare several sample sizes and number of acceptances, choose Comparison among User-defined acceptance sampling plan. The result will display sampling inspection plan using all the combinations of sample sizes and number of acceptances.

Select Measurement Type and Unit of Quality Level, and set a value.

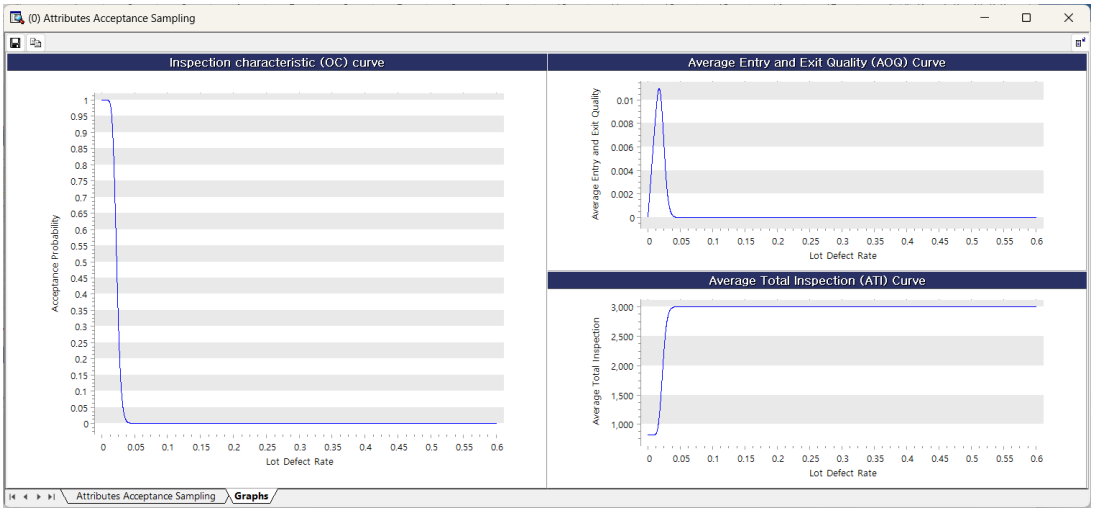


Results

An optimal sampling inspection plan is suggested based on given criteria.



Select Graphs bottom, OC curve, AOQ, and ATI charts are displayed.



---

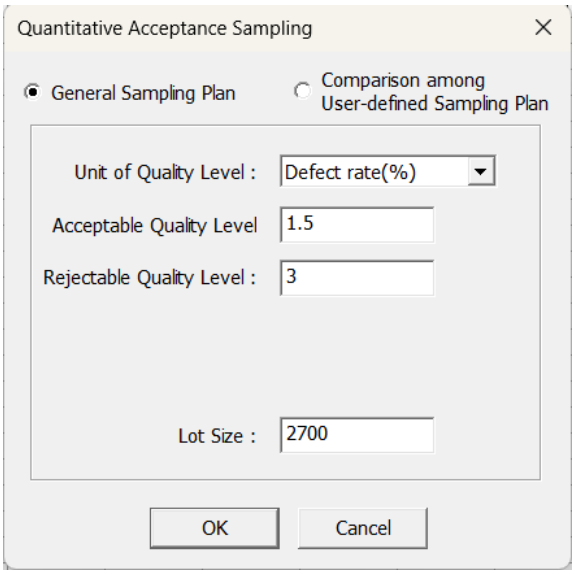
(2) Quantitative Acceptance Sampling

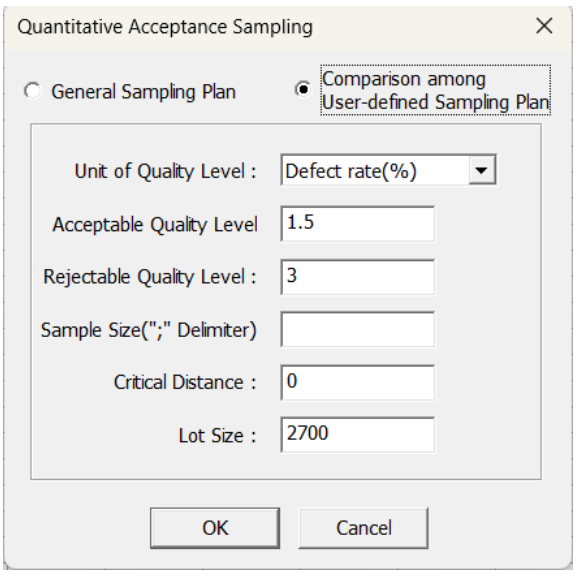
How to run

[Analyze] – [SPC] – [Acceptance Sampling] – [Quantitative Acceptance Sampling].

Select General Sampling Plan. The result gives you the optimal sample size and critical distance. If you want to compare several sample sizes with specific critical distance, choose Comparison among User-defined quantitative sampling plan. The result will display sampling inspection plan using all the combinations of sample sizes and critical distance.

Fill all the blanks and define Unit of Quality Level.

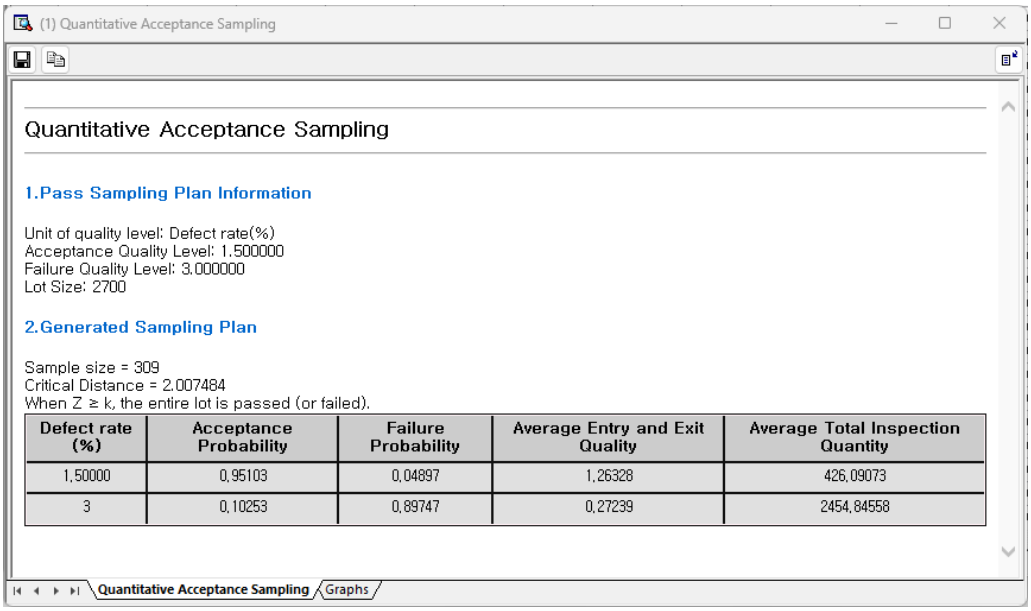




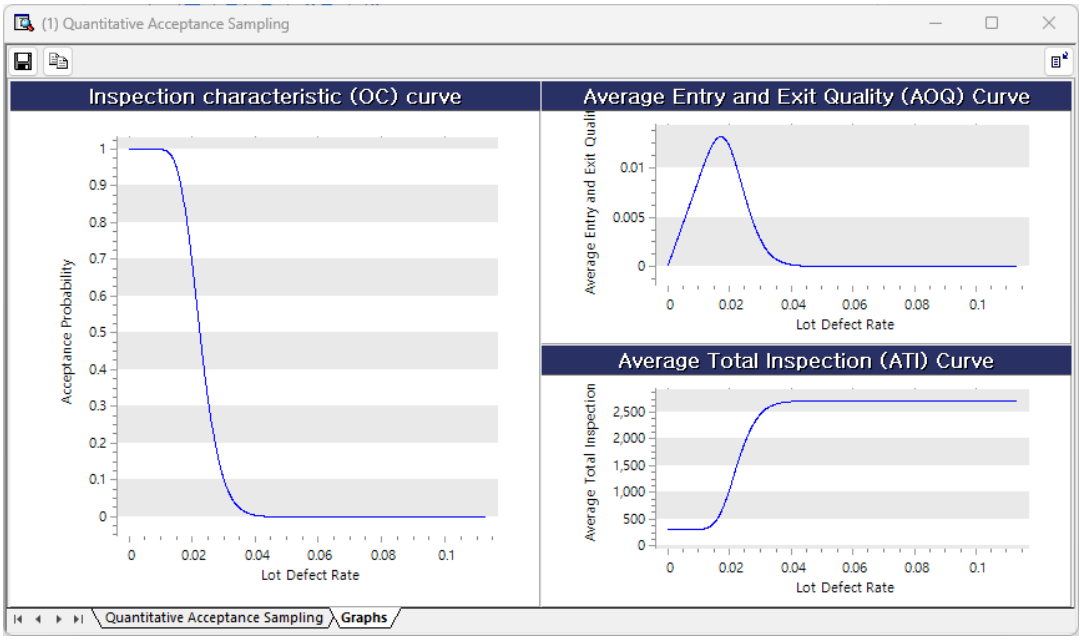
Results

The Quantitative Acceptance Sampling results is as follows.

An optimal sampling inspection plan is given based on given criteria.



Select Graphs bottom, OC curve, AOQ, and ATI charts are displayed.

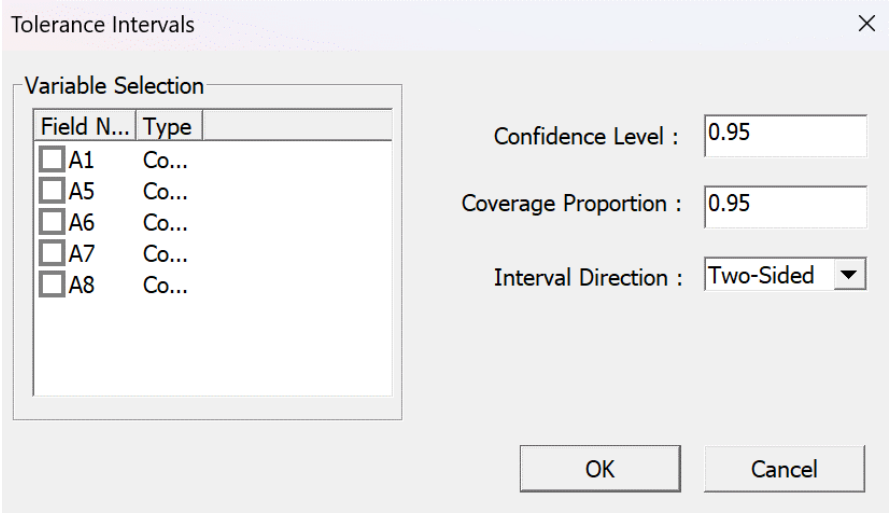


4.3.5.4 Tolerance Intervals

How to run

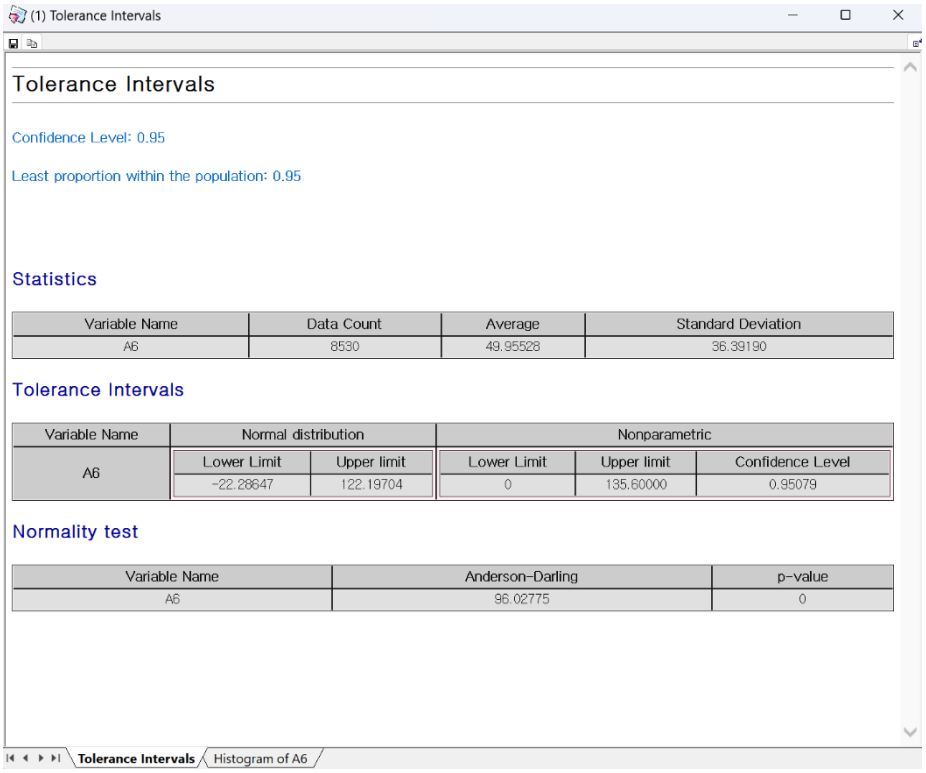
[Analyze] – [SPC] – [Tolerance Intervals]

Select the variable to get Tolerance Intervals from the Variable Selection list (multiple selections are possible), and set the Confidence Level and Coverage Proportion (0 to 1 value). Finally, select the Interval Direction (Both Side/Upper Limit/Lower Limit).

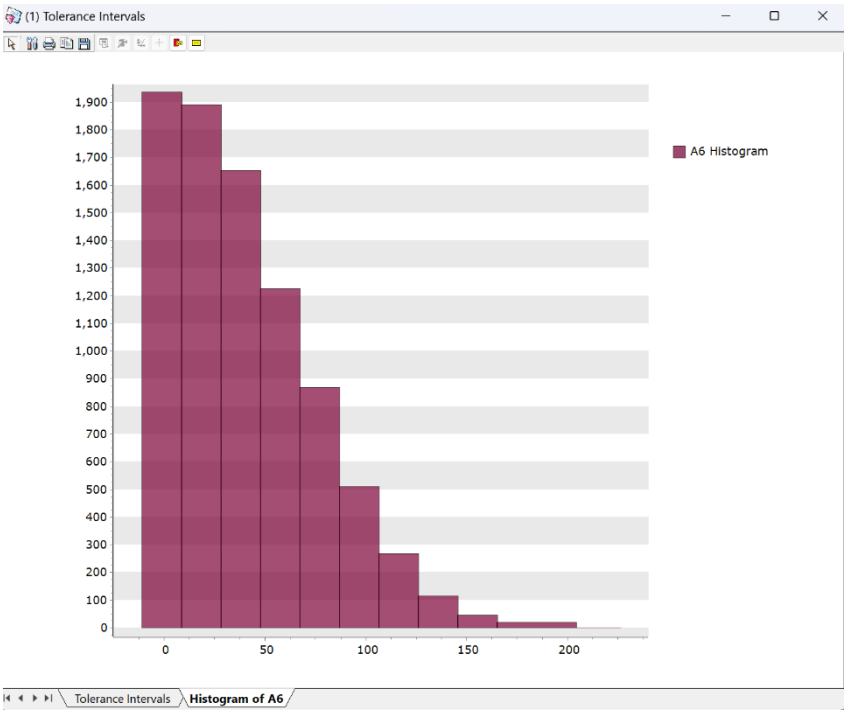


Results

- **Statistics**: average, and standard deviation.
- **Tolerance Intervals**: Satisfy the specified Confidence Level and **Coverage Proportion** from the selected variable are calculated and displayed using normal distribution or non-parametric methods.
- **Normality Test**: Use the Anderson-Darling statistics test whether the selected variable follows a normal distribution. If the p-value is closed to 1, the variable follows a normal distribution.



Histogram: A chart for distribution of the selected variable.



4.3.6 Time Series Analysis

4.3.6.1 Time Series Models

(1) Time Series Decomposition

Overview

Time series decomposition means dividing the data into components. Specifically, time series data can have components such as trend and seasonality, and these components can be extracted and separated. Classical decomposition includes multiplicative and additive decomposition.

The structure of classical decomposition methods is as follows.

Multiplicative Decomposition
Additive Decomposition

The statistics obtained through decomposition are summarized as follows.

Trend Data$(\mathrm{tr_t})$
Detrended Data
Seasonal Index Data$(\mathrm{S_t})$
Deseasonal Index Data$(\mathrm{y_t - S_t})$
Fitted Value$(\hat{y}_t)$
Residual

Multiplicative Decomposition (Including Trend)

The Multiplicative Decomposition Model has the following form.

$$\mathrm{y_t = TR_t \times SN_t \times CL_t \times IR_t}$$

yt: Observed value at time t.

TRt: Trend component at time t.

SNt: Seasonal component at time t.

CLt: Cyclical component at time t.

IRt: Irregular component at time t.

Additive Decomposition (Including Trend)

The Additive Decomposition Model has the following form.

$$y_t = TR_t + SN_t + CL_t + IR_t$$

yt: Observed value at time t.

TRt: Trend component at time t.

SNt: Seasonal component at time t.

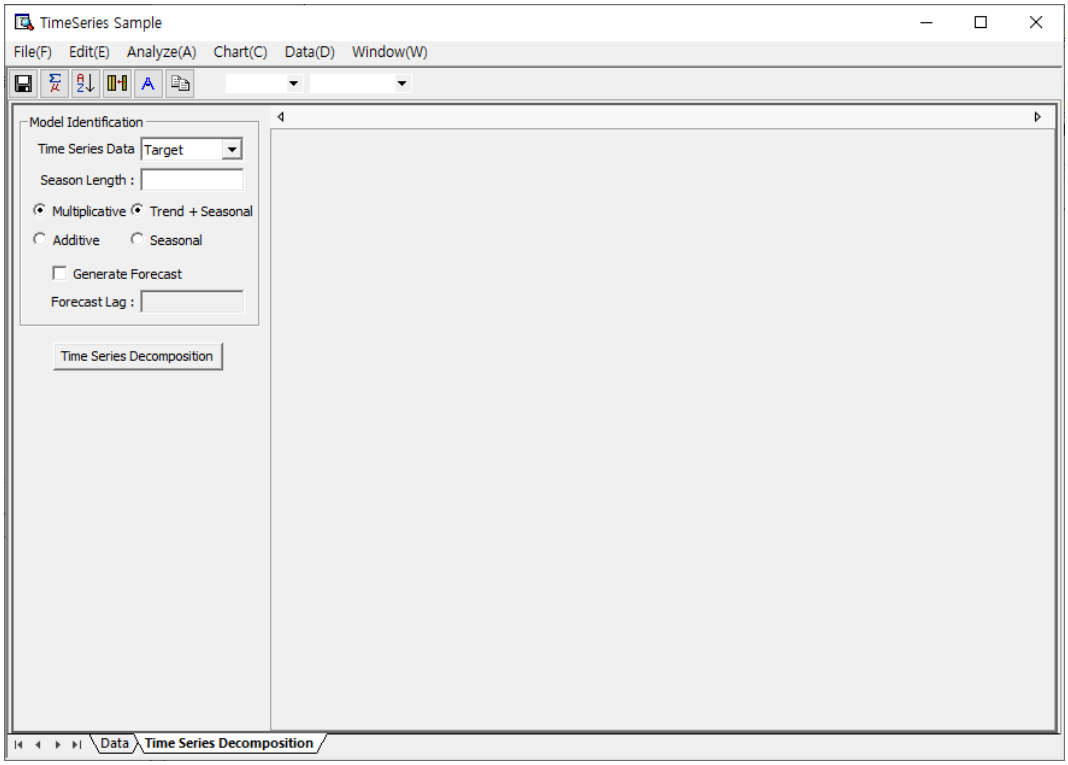
CLt: Cyclical component at time t.

IRt: Irregular component at time t.

At this point, the trend can be excluded from the analysis.

How to run

[Analyze] – [Time Series Analysis] – [Time Series Models] – [Time Series Decomposition]



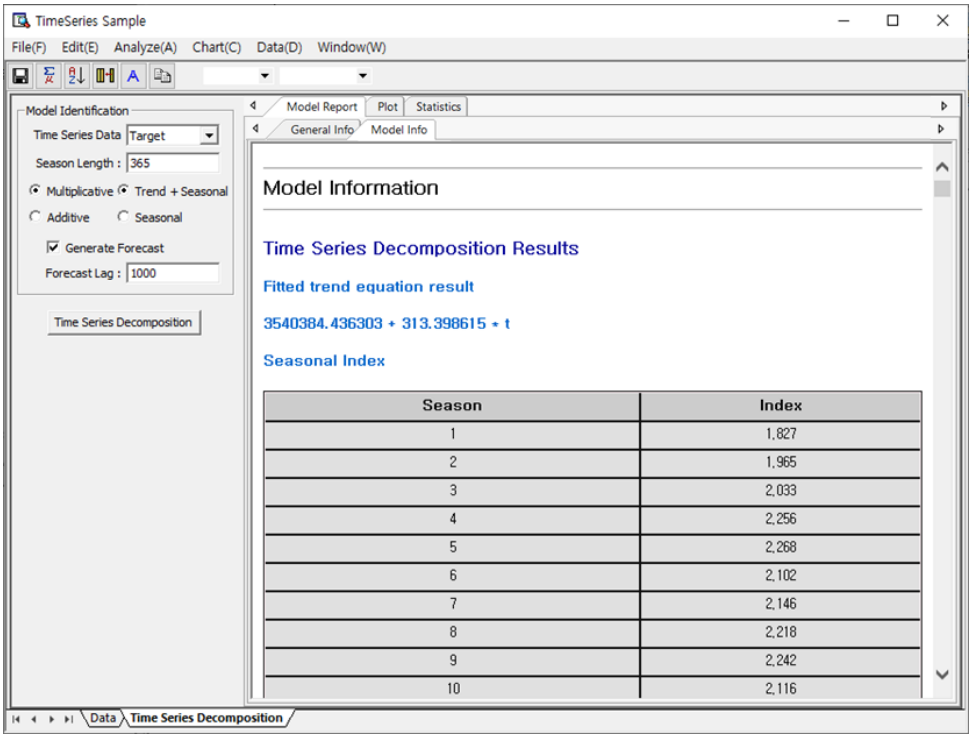
Select time series data on the main screen, select the length of the season, and whether to use Multiplicative Decomposition or Additive Decomposition, and then select either Trend+Seasonal or Seasonal. If you want a forecast, click Generate Forecast and enter Forecast Lag.

Results

- **Model Report**

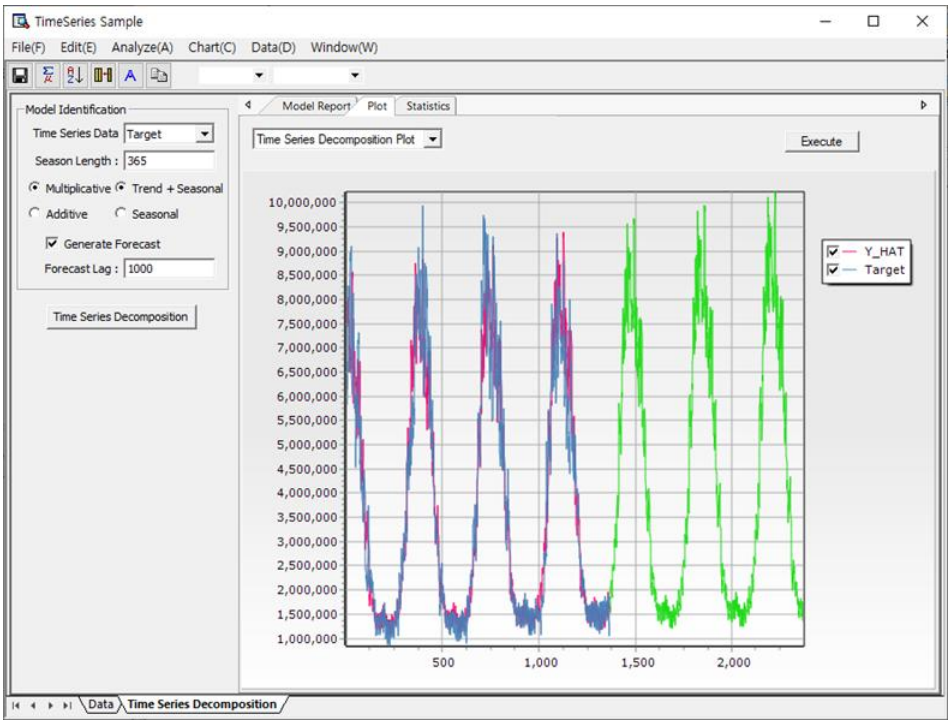
  General Info: Shows basic information about time series data.

  Model Info: The forecast results are displayed when the trend and/or seasonality are selected in the Generate Forecast.

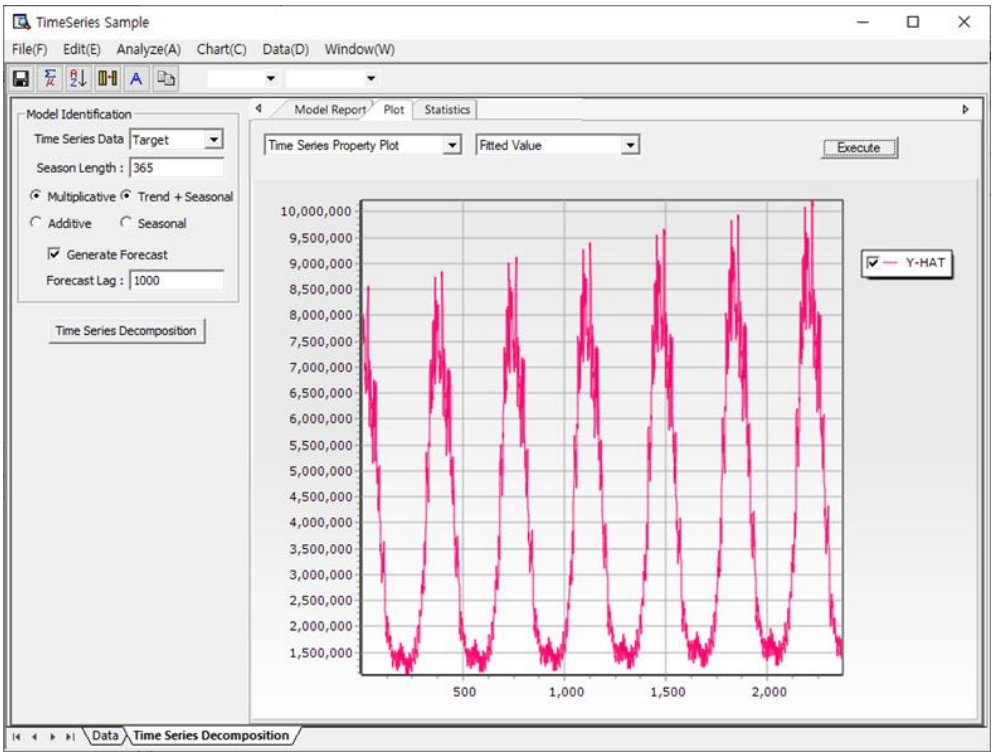


- **Plot**

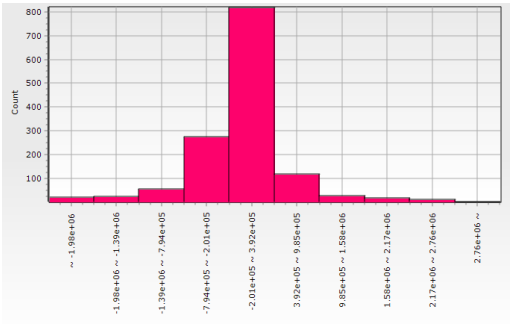
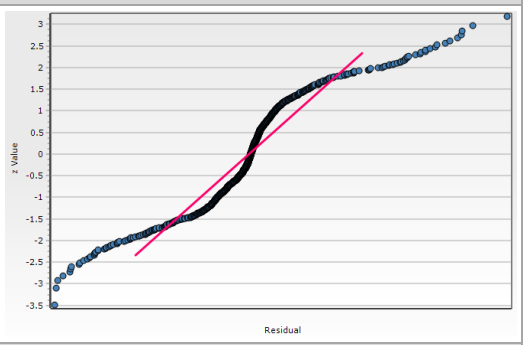
This function is provided for visual interpretation of data obtained through **Time Series Decomposition**. **Time Series Decomposition** provides **Decomposition results**, time series property and **Residual** plots.

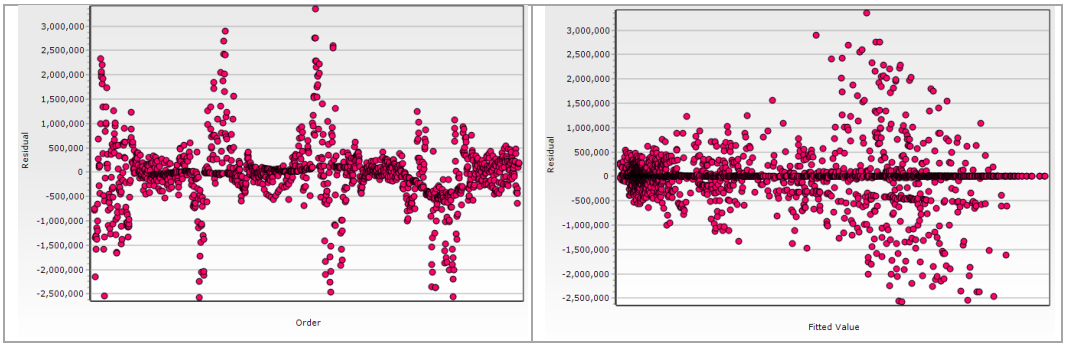


Time Series Decomposition Plot shows **Original Data, Fitted Data**.



The **Time Series Property Plot** shows each decomposed component as a plot. You can view all data at once, or view graphs for each piece of data separately.

Histogram	Normal Probability Plot
	
Residual vs. Order	**Residual vs. Fitted Values**



Through Residual Plot (Histogram, Normal Probability Plot, Residual vs. Order, Residual vs. Fitted Values), you can analyze the residuals obtained as a result of decomposition.

- **Statistics**

  Data obtained through **Time Series Decomposition** analysis is displayed in table form. It also provides a function to save it.

---

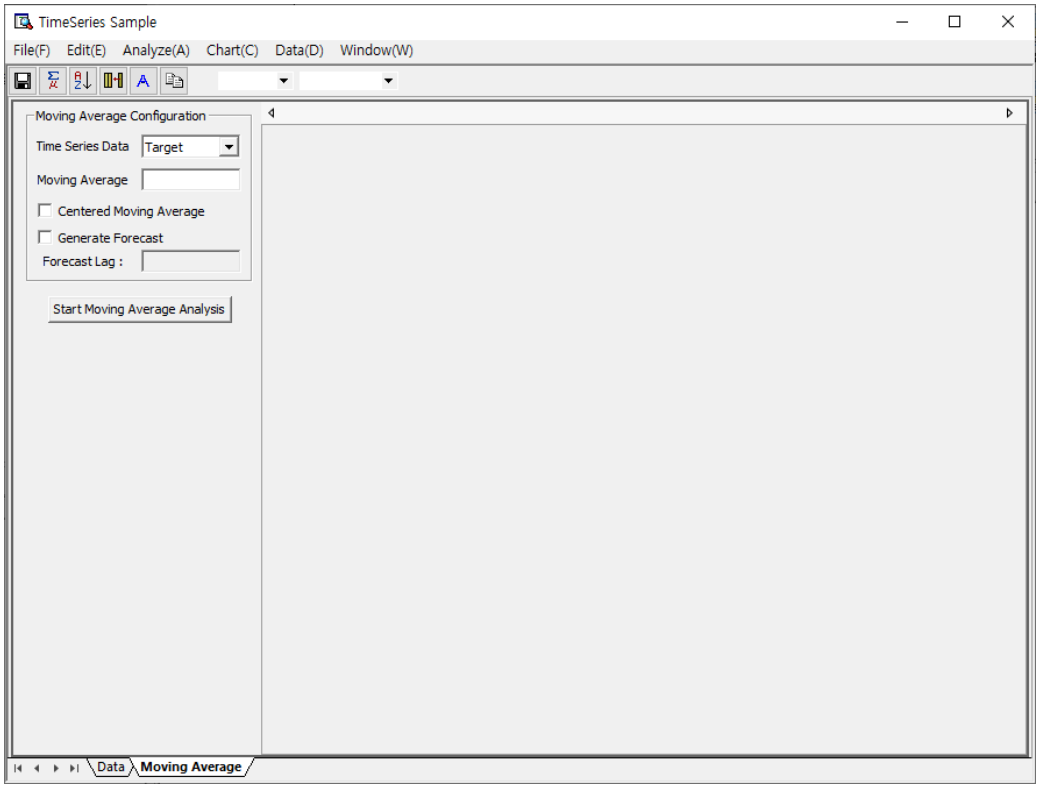
(2) Moving Average

Overview

The Moving Average method refers to the average value of several data from the present, past, and future. The method is a widely used one in its simplest form along with its expanded Centered Moving Average.

How to run

[Analyze] – [Time Series Analysis] – [Time Series Models] – [Moving Average]



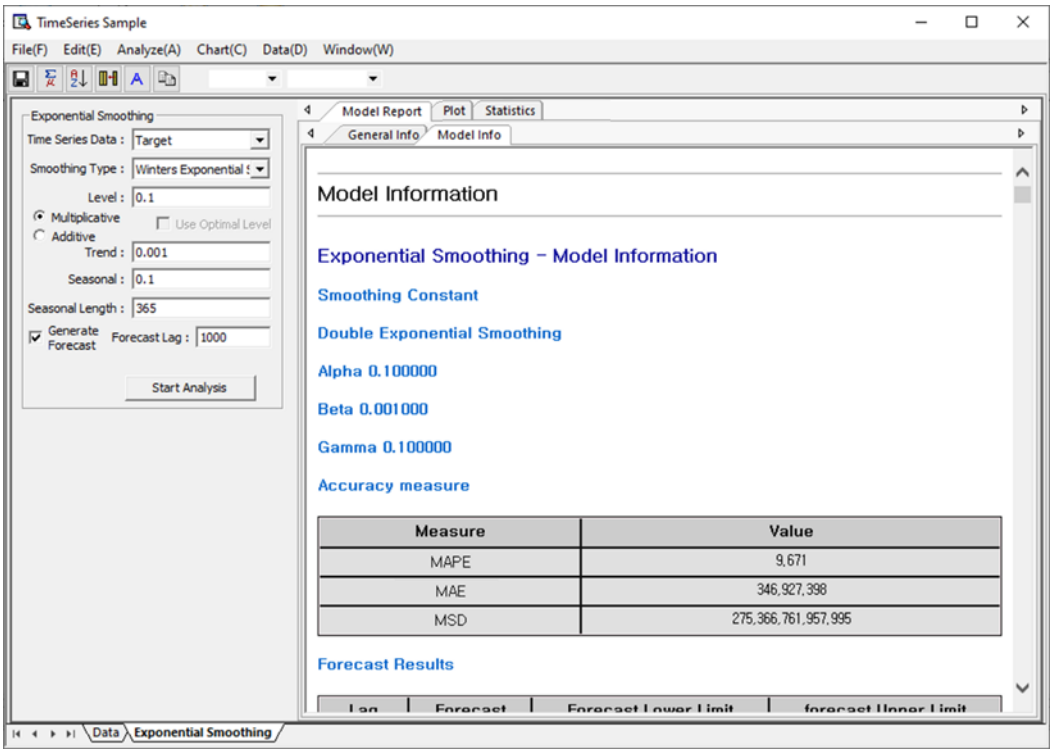
Select Time Series Data on the main screen and enter the length of the Moving Average, Centered Moving Average option, and Forecast Lag if you want to make a forecast. After completing the input, click the [Start Moving Average Analysis] button.

Results

- **Model Report**

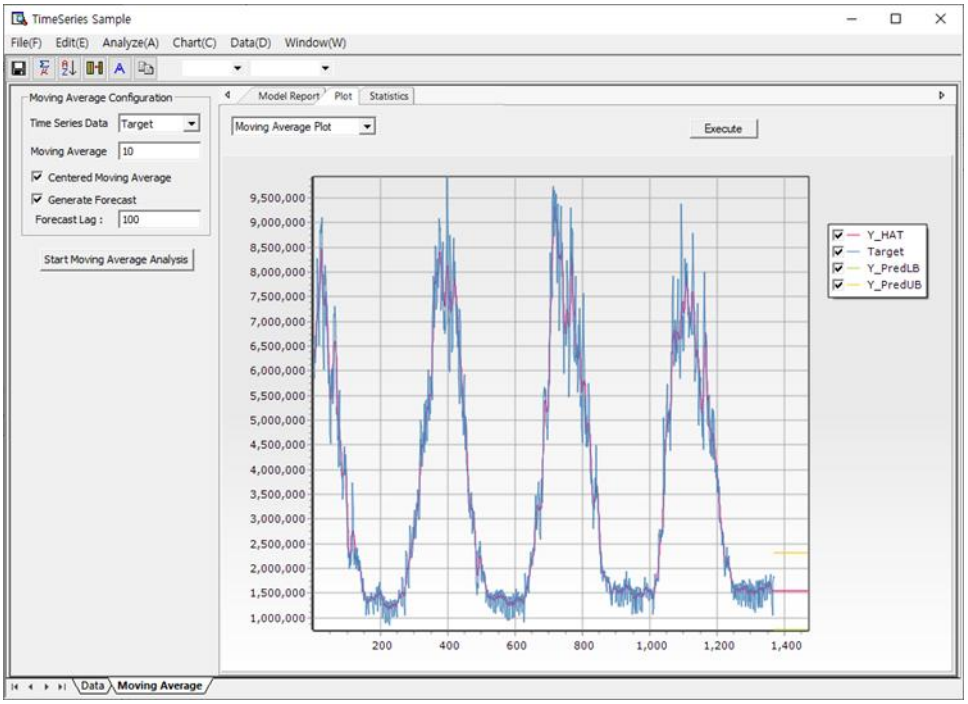
  General Info: Provides basic information about time series data.

  Model Info: Provides information on the results of the Moving Average analysis. You can also view the results of the prediction accuracy based on MAPE, MAE and MSD.



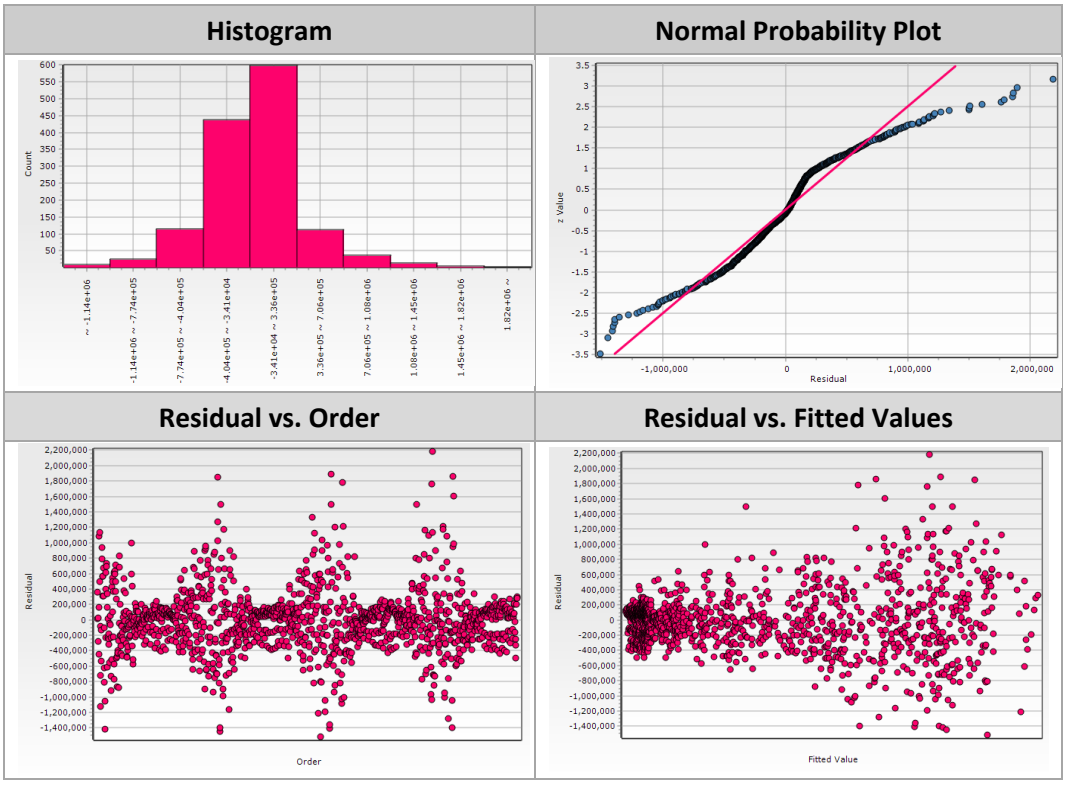
- **Plot**

You can visually view the data obtained as a result of Moving Average.



Moving Average plot shows time series data, fitted values, and prediction-related statistics.

You can get the results the following Residual Plots (Histogram, Normal Probability Plot, Residual vs. Order, Residual vs. Fitted Values).



- **Statistics**

  Data obtained through Moving Average analysis is displayed in table form. It also provides a function to save it.

---

(3) Exponential Smoothing

Overview

- **Single Exponential Smoothing**

  The following model has a constant term without any trend,

$$y = \beta_0 + \epsilon_t$$

  Using the least squares method, we have the estimate for the constant parameter.

$$\hat{\beta}_0 = \bar{y} = \sum_{t=1}^{n} \frac{y_t}{n}$$

If you look at the above equation, you can see that in the process of calculating the estimate for β_0, the same size weight (1/n) is assigned to all observations, but this does not seem reasonable. So the idea of exponential smoothing is to give more weight to recent data and less weight to older data. Among these exponential smoothing methodologies, the one that processes data without trends and seasonality is **Single Exponential Smoothing**.

- **Double Exponential Smoothing**

  Let's consider the following time series model.

$$y_t = \beta_0 + \beta_1 t + \epsilon_t$$

  If β_0, β_1 are the parameters with fixed values, the parameters can be estimated by the least squares estimation, for example. On the other hand, if β_0, β_1 can change over time, **Double Exponential Smoothing** is proposed to explain this model.

- **Winters' Method: Multiplicative Winters' Method**

  The Multiplicative Winters' method is a time series model that considers seasonality as well as level and trend. The Multiplicative Winters' Method is known to be suitable for forecasting of time series expressed by the following equation.

$$y_t = (\beta_0 + \beta_1 t) * SN_t + \epsilon_t$$

- **Winters' Method: Additive Winters' Method**

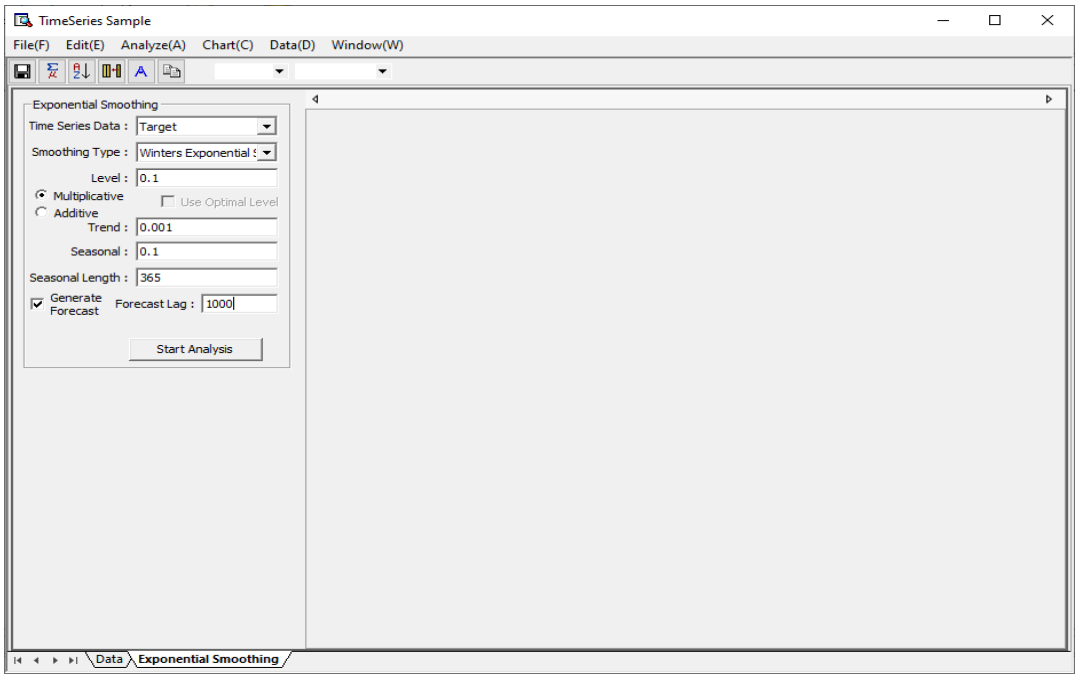
  The Additive Winters' method is known to be most suitable for forecasting of time series data that satisfies the following equation.

$$y_t = (\beta_0 + \beta_1 t) + SN_t + \epsilon_t$$

The Additive Winters's Method can be obtained by slightly modifying the Multiplicative one.

How to run

[Analyze] – [Time Series Analysis] – [Time Series Models] – [Exponential Smoothing]



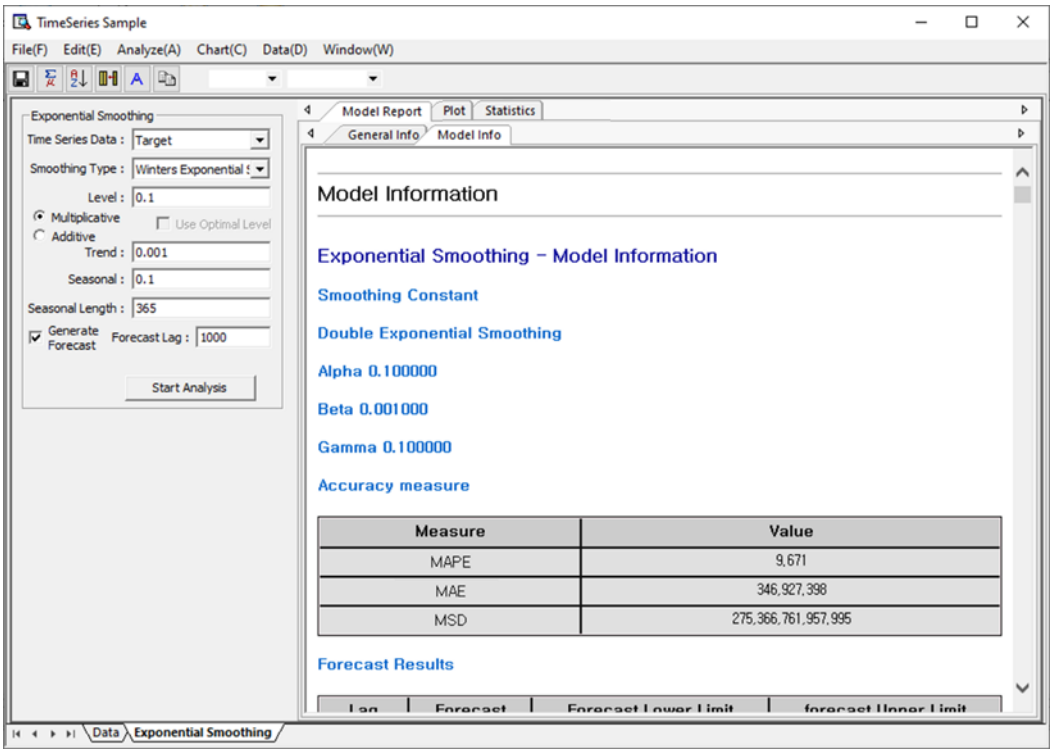
On the main screen, select Time Series Data and select which smoothing type to use. At this time, when using Single Exponential Smoothing, the user can enter the level value directly, or check 'Use Optimal Level' to automatically enter the optimal level value between 0 and 1. When selecting Double Exponential Smoothing, enter level and trend smoothing constants. When using the Winters' Method, select multiplicative and additive and enter level, trend, seasonal smoothing constants, and seasonal length. If you want to make a prediction, please enter Forecast Lag.

Results

- **Model Report**

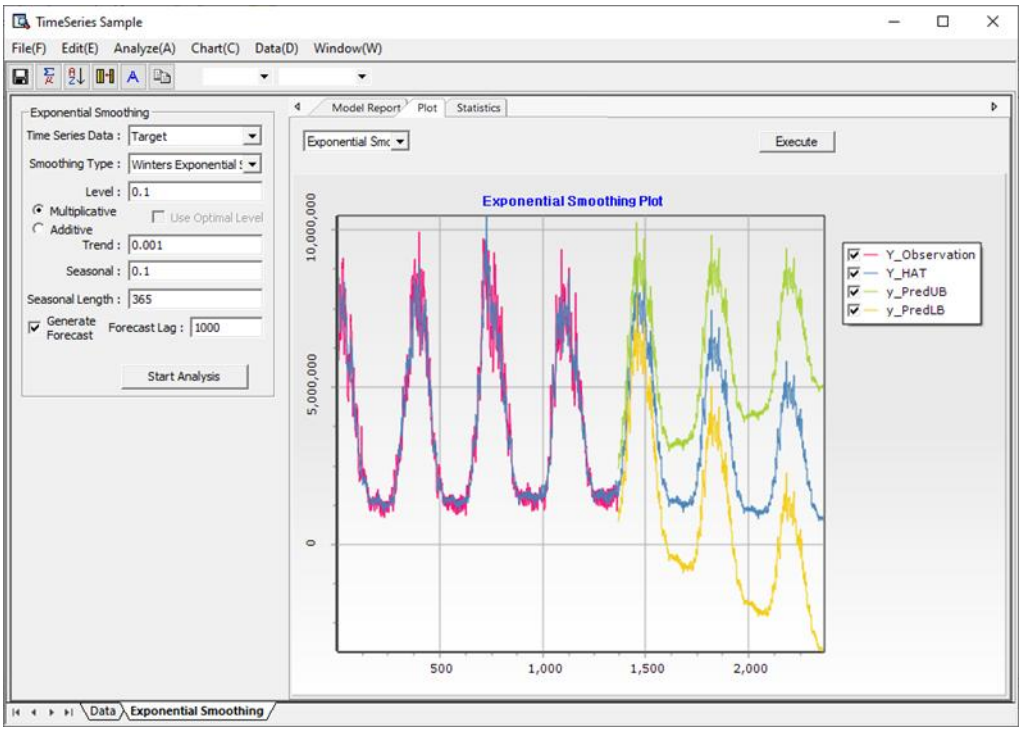
  General Info: Shows basic information about time series data.

  Model Info: Provides information obtained through exponential smoothing analysis. You can also view the results of the prediction accuracy based on MAPE, MAE and MSD.

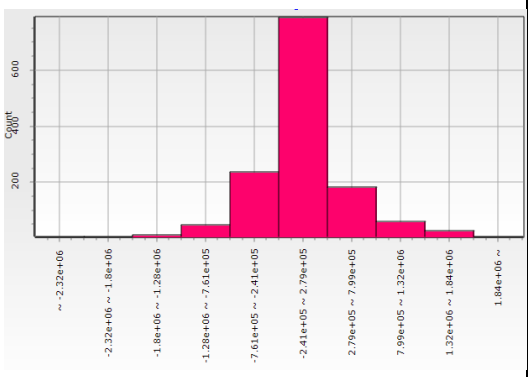
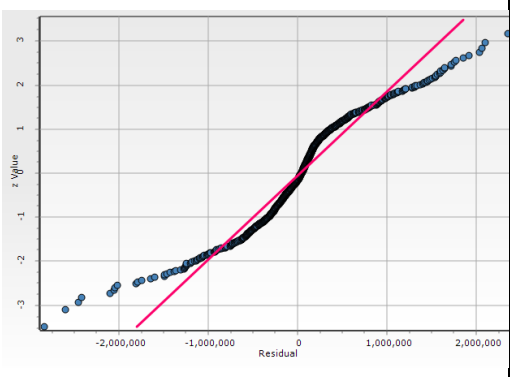
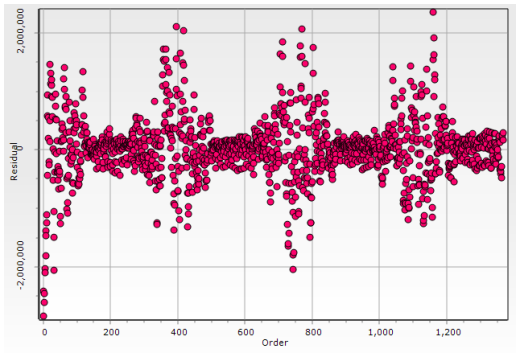
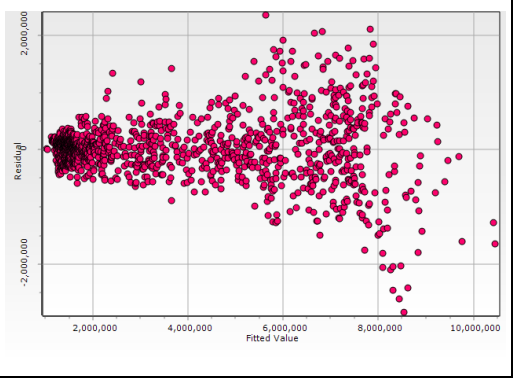


- **Plot**

You can visually view the data obtained as a result of Exponential Smoothing analysis.



The Exponential Smoothing Plot allows you to view the time series data, fitted values, and the forecast-related values.

Histogram	Normal Probability Plot
	
Residual vs. Order	**Residual vs. Fitted Values**
	

You can get the results the following Residual Plots (Histogram, Normal Probability Plot, Residual vs. Order, Residual vs. Fitted Values) in the Exponential Smoothing analysis.

- **Statistics**

  The statistics obtained as a result of the Exponential Smoothing analysis can be seen in the table. It also provides the function to save it.

---

(4) ARIMA

Overview

The ARIMA model is the most commonly used model in Univariate Time Series Analysis and satisfies the following equation. In case of ARIMA(p, d, q),

$$\phi(B)(1-B)^d z_t = \theta(B)a_t \text{ where } a_t \text{ is i.i.d normal white noise}$$

$$(1-\phi_1 B-\phi_2 B^2-\cdots-\phi_p B^p)(1-B)^d z_t = (1-\theta_1 B-\theta_2 B^2-\cdots-\theta_q B^q)a_t$$

$$(1-\varphi_1 B-\varphi_2 B^2-\cdots-\varphi_{p+d} B^{p+d})z_t = (1-\theta_1 B-\theta_2 B^2-\cdots-\theta_q B^q)a_t$$

the purpose is to estimate the coefficients from the above equation, test the suitability of the equation created using the estimated coefficients, and finally perform forecasting.

- **Box-Jenkins' Method**

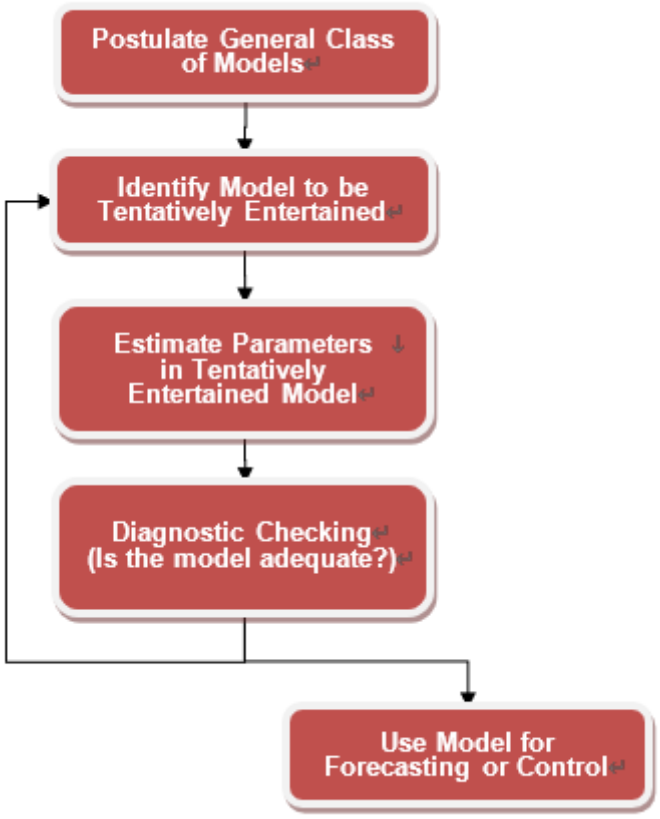
Box and Jenkins explained the processes of Time Series Analysis in the following ways.

1. From the interaction of theory and practice, a useful class of models for the purposes at hand is considered.

2. Because this class is too expensive to be conveniently fitted directly to data, rough methods for identifying subclass of these models are developed. Such methods of model identification employ data and knowledge of the system to suggest an appropriate parsimonious subclass of models which may tentatively entertained. In addition, the identification process can be used to yield rough preliminary estimates of the parameters in the model.

3. The tentatively entertained model is fitted to the data and the parameters in the model are to be estimated. The estimates obtained during the identification stage can now be used as initial values in more refined iterative methods for estimating the parameters.

4. Diagnostic checks are applied with the object of uncovering possible lack of fit and diagnosing the cause. If no lack of fit is indicated, the model is ready to use. If any inadequacy is found, the iterative cycle of identification, estimation, and diagnostic checking is repeated until a suitable representation is found.



To put it simply,

1. Assume a general model (ARIMA, ARCH, for example).

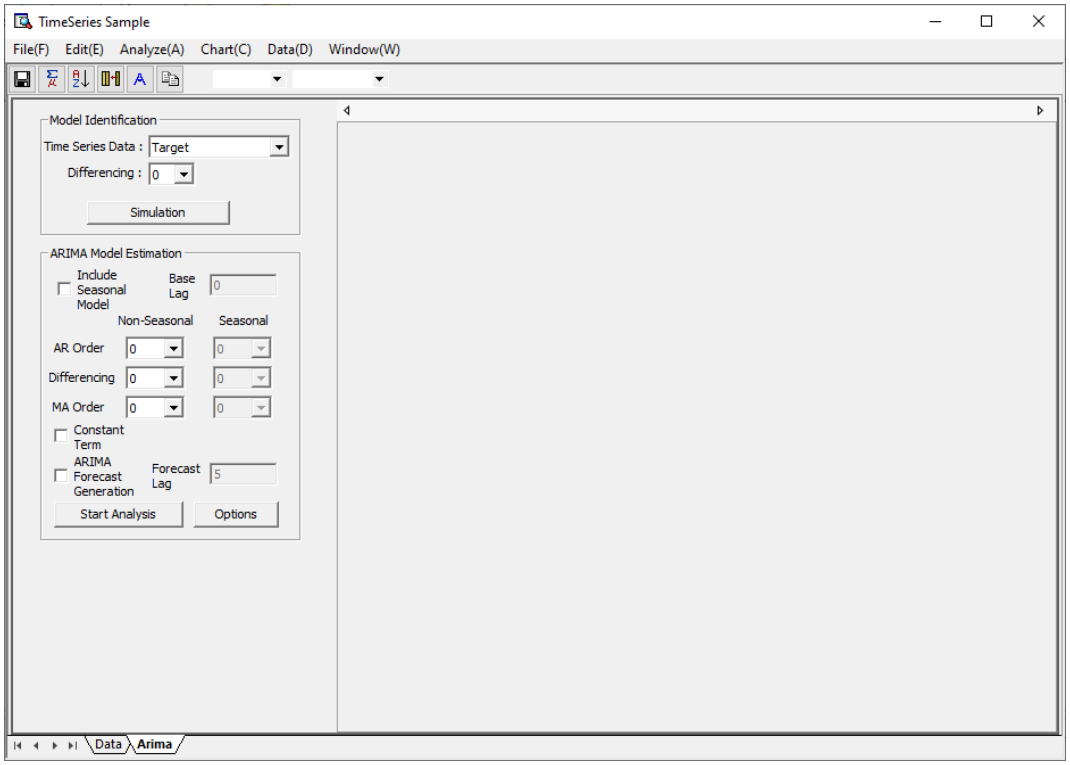
2. Determine a candidate model (such as ARIMA(1,1,1)) from the models.

3. Estimate the parameters of the model.

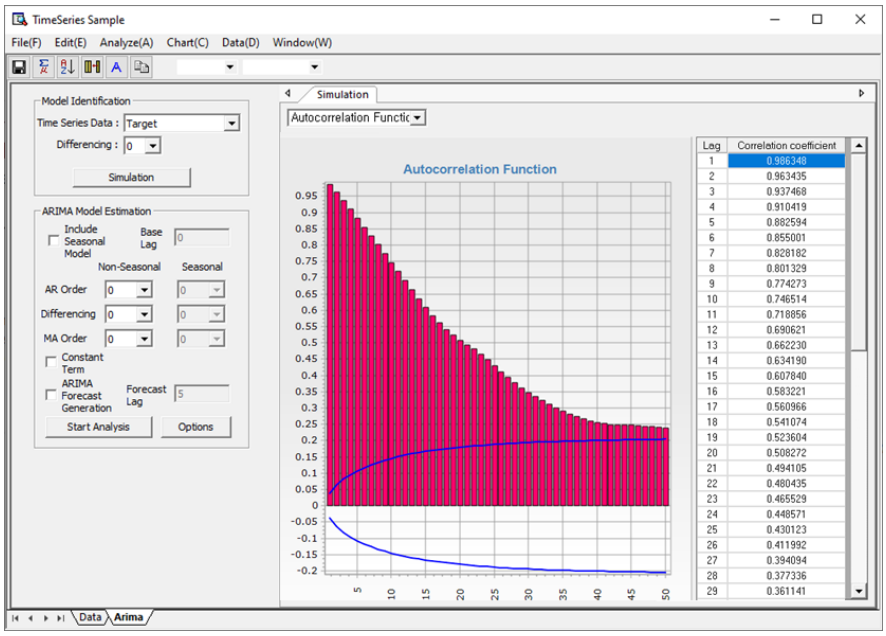
4. Verify suitability, and further work such as forecasting is carried out.

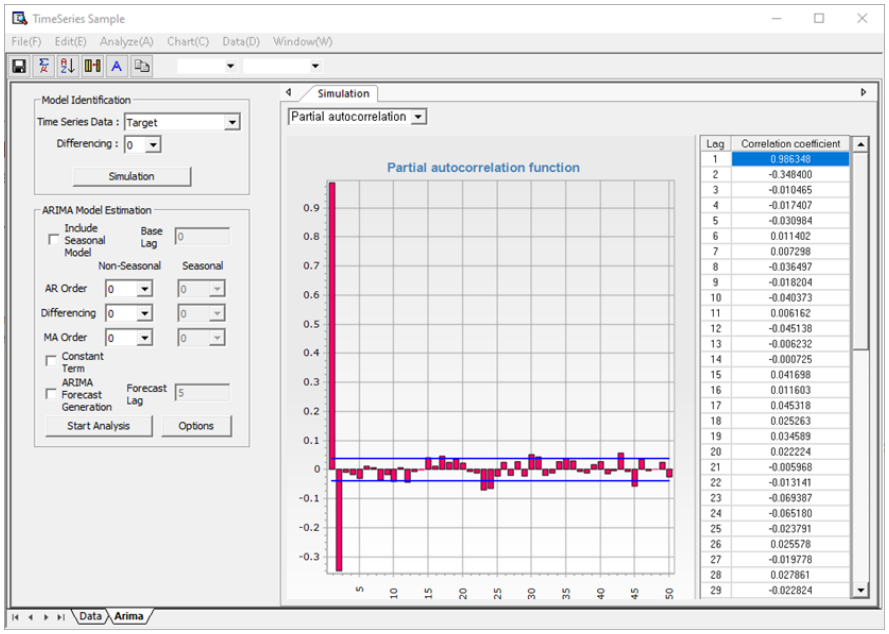
How to run

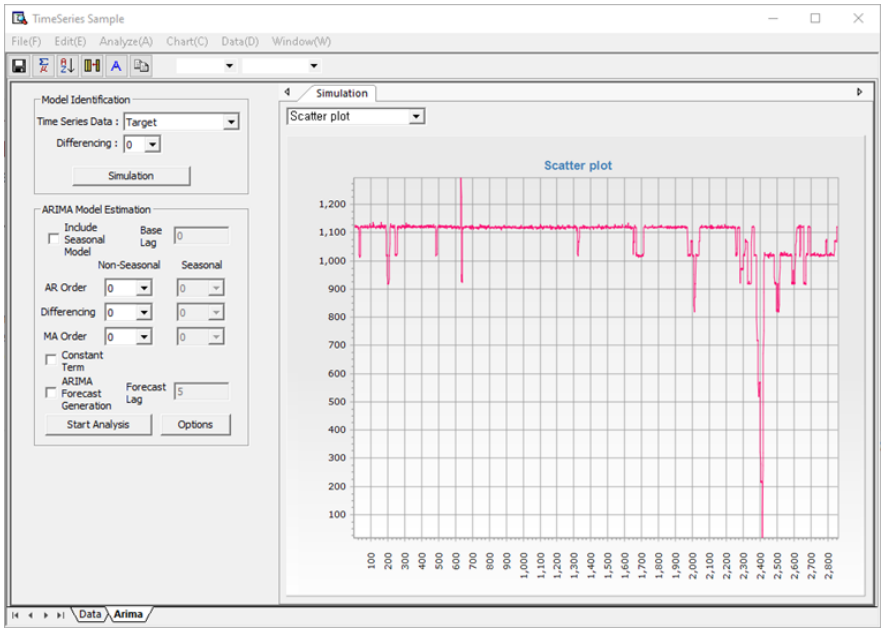
[Analyze] – [Time Series Analysis] – [Time Series Models] – [ARIMA]



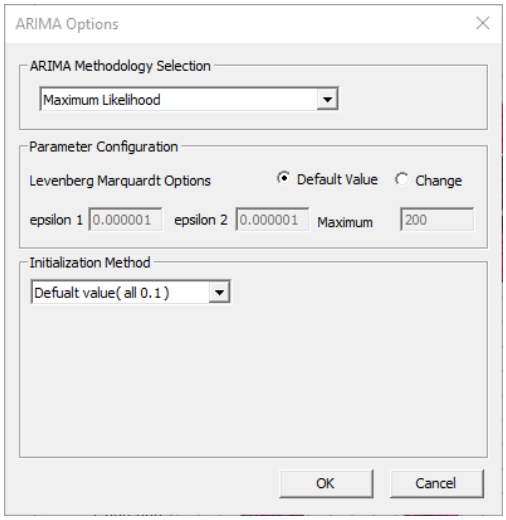
On the ARIMA main screen, you can run the Model Identification process and ARIMA Model Estimation process. Through Model Identification, you can see the autocorrelation, partial autocorrelation, and scatter plot of the differential time series data, which can help users determine the orders of the ARIMA model needed when estimating the ARIMA model. Below are examples of autocorrelation, partial autocorrelation, and scatter plot.







In this way, models are identified and orders are determined through Autocorrelation, Partial Autocorrelation, and Scatter Plot. If you click the Options button, you can select the ARIMA methodology (Conditional Least Square, Maximum Likelihood) as follows and enter constant settings related to Levenberg Marquardt used for Parameter Optimization, and initialization method.



And when you start Time Series Analysis, you can get the following results:

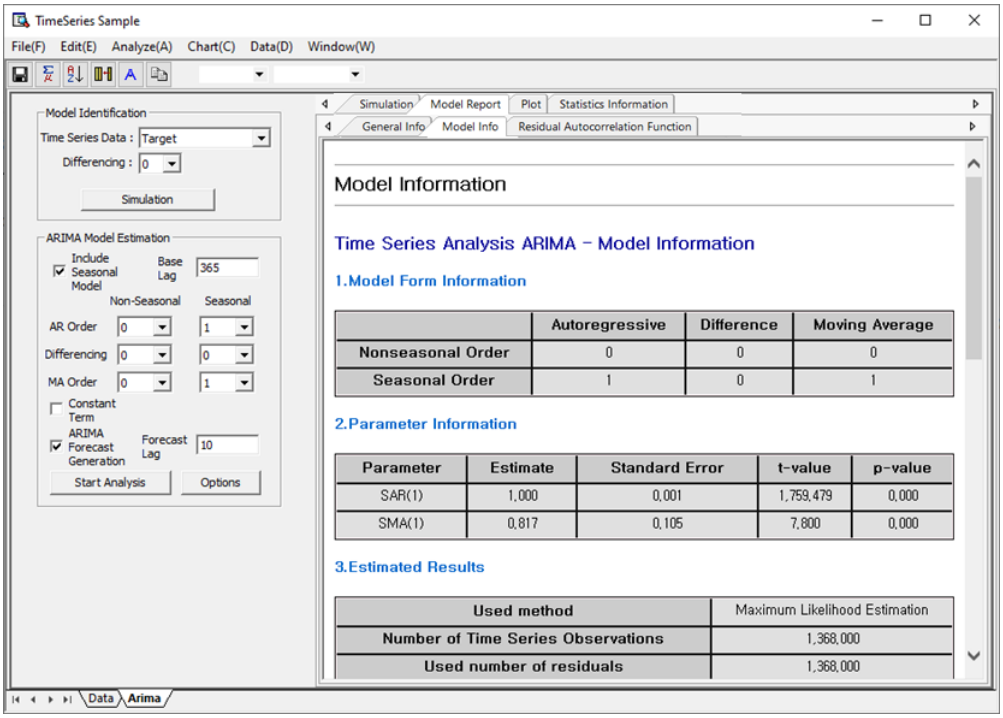
Results

- **Model Report**

  General Info: Shows basic information about time series data.

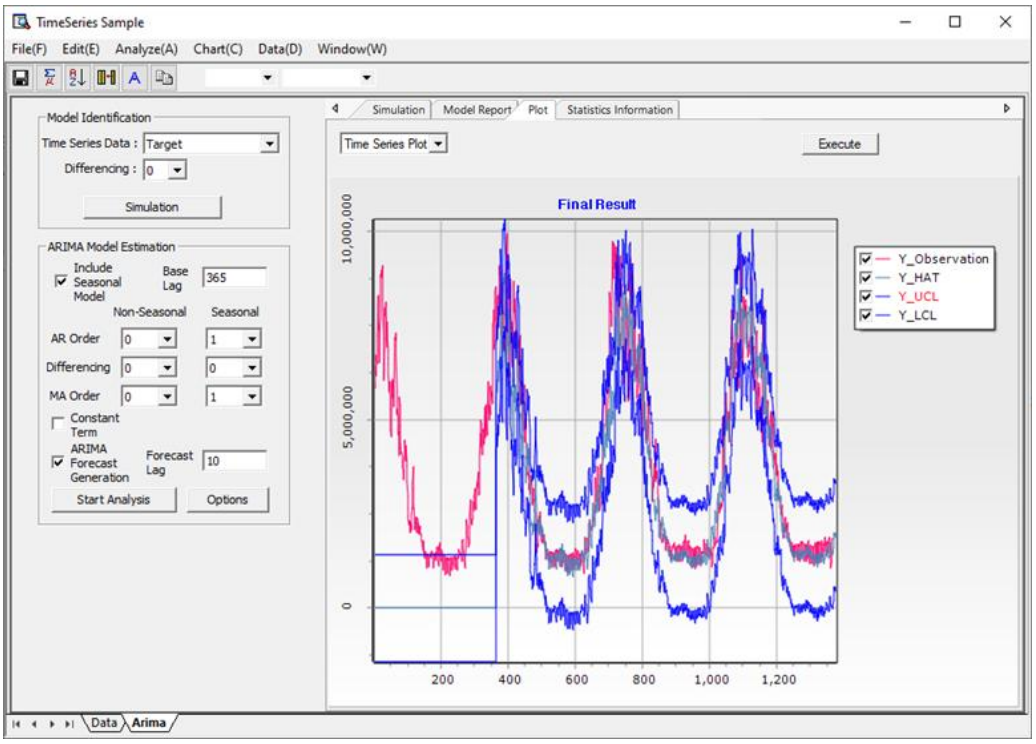
  Model Info: Provides the results of **Time Series Analysis**. You can obtain model shape information, parameter information, Parameter Optimization results, and forecast results.

  Residual Autocorrelation Function: You can perform a diagnostic check through the Residual Autocorrelation Function and Residual Partial Autocorrelation Function.

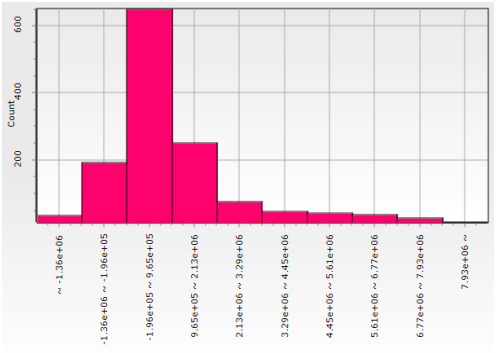
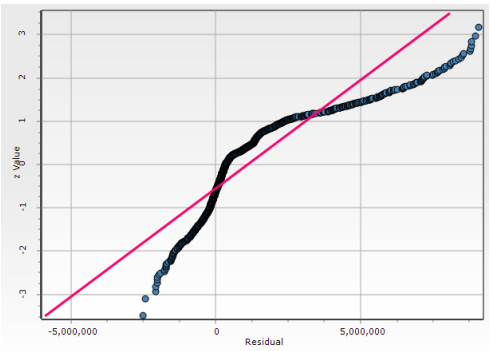
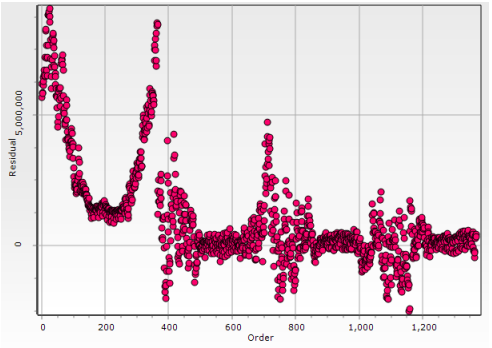
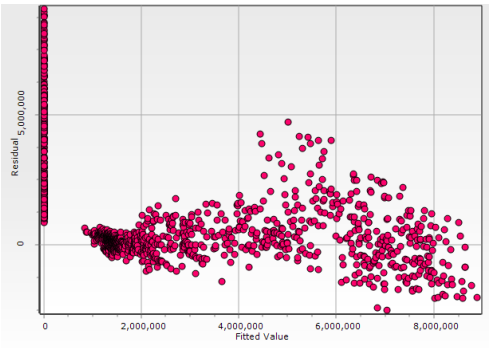


- **Plot**

Data obtained through ARIMA models can be visually analyzed. Through a time series plot, you can see observed values, fitted values, predicted values, and upper and lower prediction limits.



You can perform residual analysis for diagnostic check through the following residual plots.

Histogram	Normal Probability Plot
	
Residual vs. Order	**Residual vs. Fitted Values**
	

- **Statistics**

  Through Statistics Information, you can view the data obtained after creating the model in a table. It also provides a function to save the information obtained in this way.

---

(5) Trend Analysis

Overview

Trend analysis is a methodology performed to obtain basic trends in time series data. Time series data observed at each point in time are created under the logic that they are explained only by functional relationships with time as the horizontal axis and observed values as the vertical axis. Let us limit these functional relationships to the following.

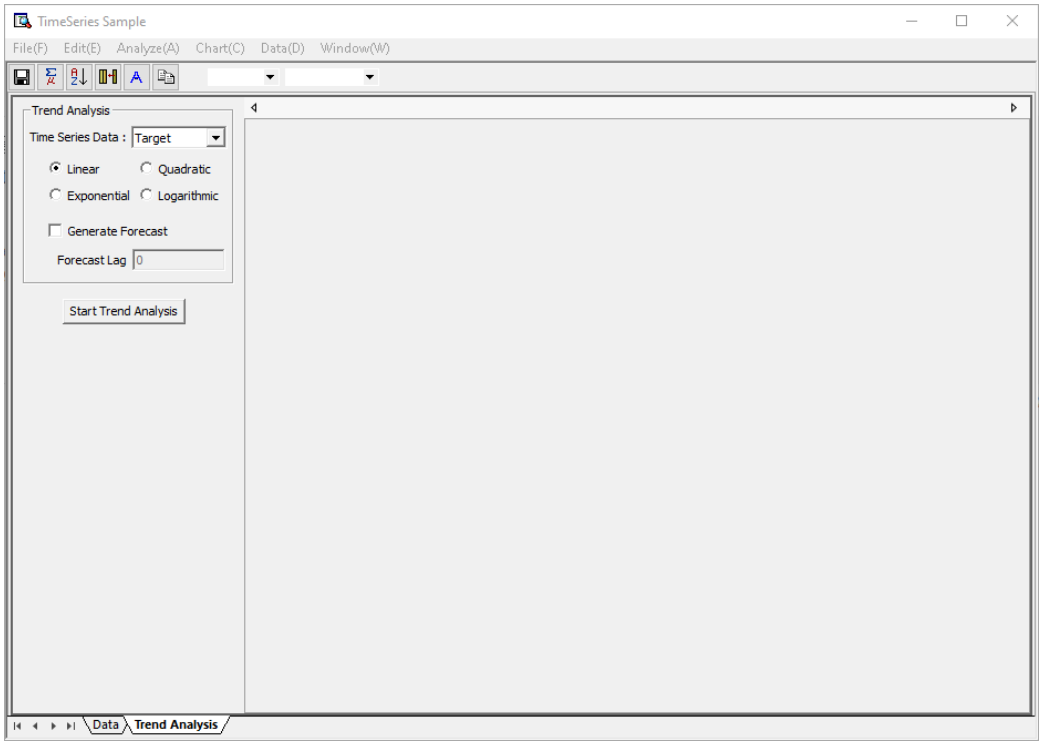
Linear Quadratic Exponential Growth Logarithmic

All parameters estimated for each function are obtained in closed form through the Least Square method.

How to run

[Analyze] – [Time Series Analysis] – [Time Series Models] – [Trend Analysis]

In the window, select what type of function to fit, and if you want to make a forecast, enter Forecast Lag and click the Start Trend Analysis button.

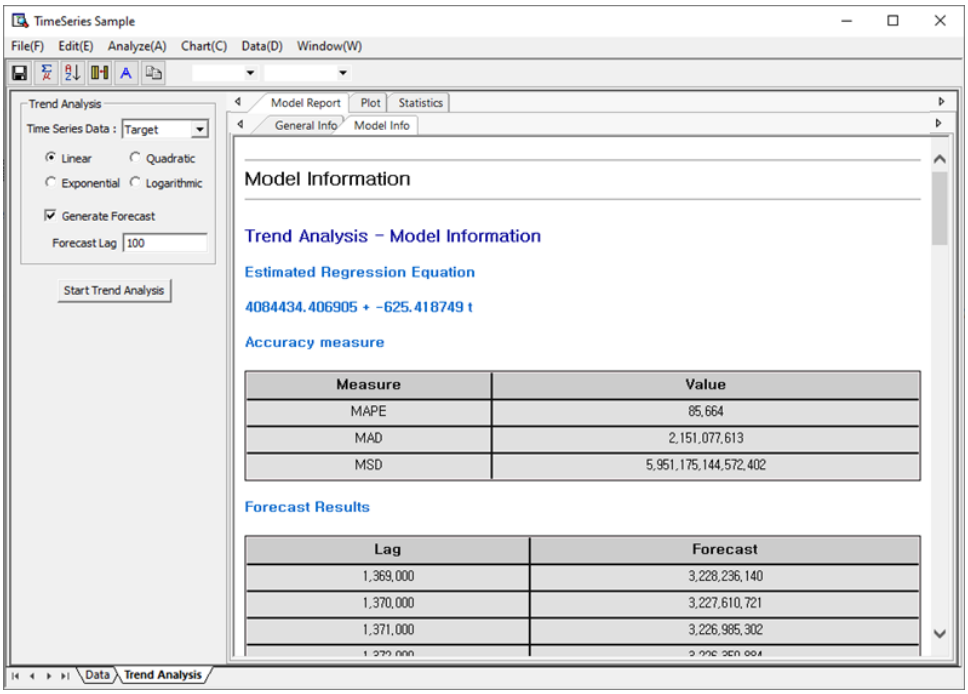


Results

- **Model Report**

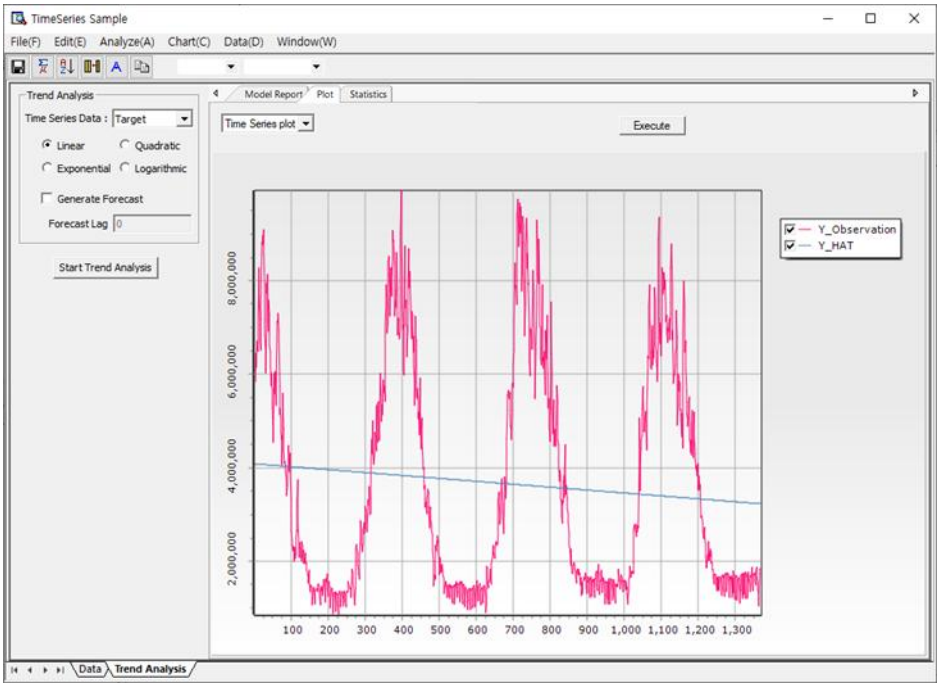
General Info: Shows basic information about time series data.

Model Info: Provides regression equations, accuracy measures, and forecast results obtained through trend analysis.

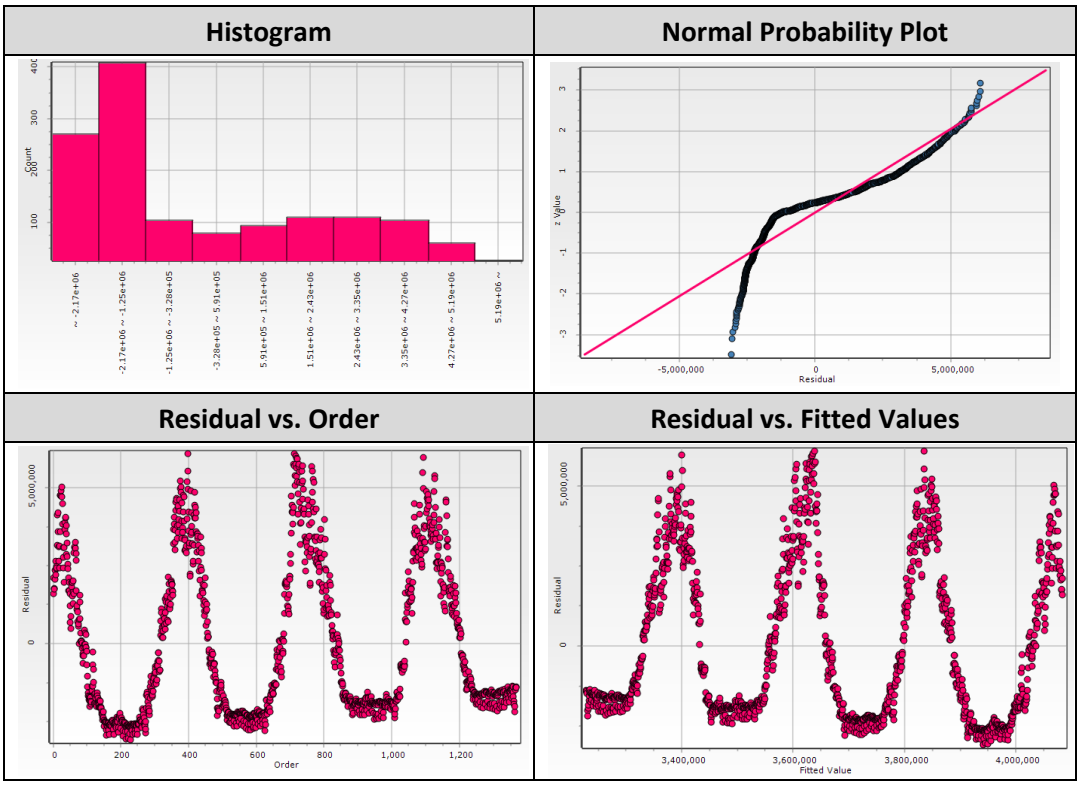


- **Plot**

Visually displays the original and fitted data obtained through trend analysis.



Through the Time Series Plot, you can see the time series data and the regression line that estimates it at a glance.



Residual Plot (Histogram, Normal Probability Plot, Residual vs. Order, Residual vs. Fitted Values) allows you to analyze residuals obtained by Trend Analysis.

- **Statistics**

  Statistics obtained through trend analysis can be viewed in the table. In addition, it also provides the function to save tables.

---

(6) GARCH

Overview

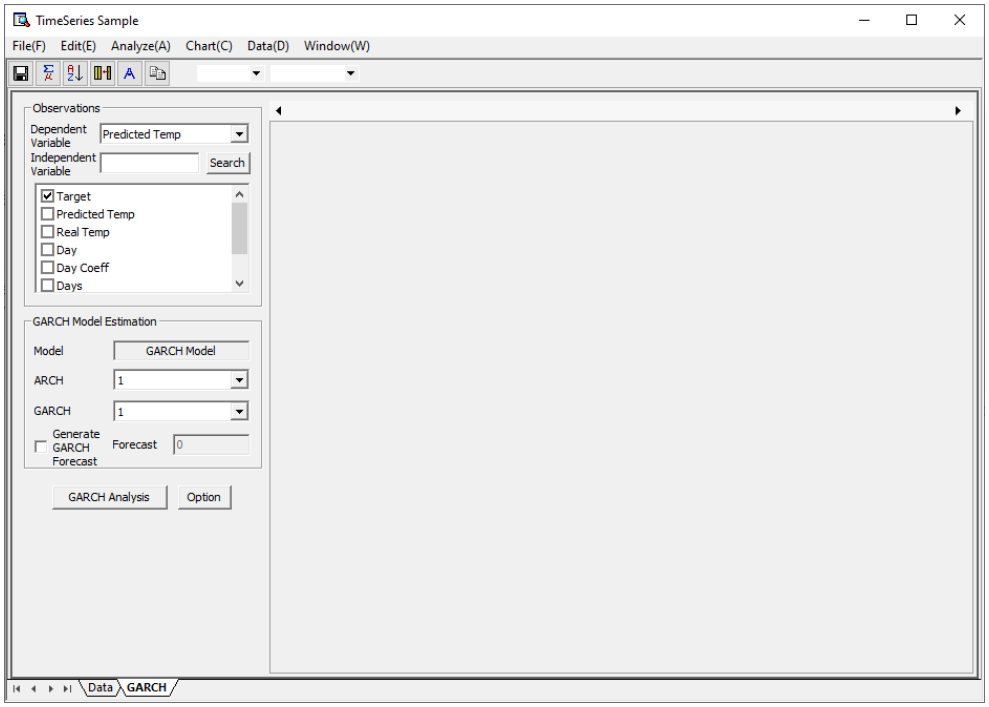
The ARCH (Autoregressive Conditional Heteroskedasticity) methodology was proposed in the 1980s, and Robert Engle, who proposed this method, received the Nobel Prize in Economics in 2003. The innovative aspect of this methodology is to modify the assumption that random shocks have constant variance in the time series analysis so far, and introduce the assumption that random shocks have constant variance unconditionally, but the conditional variance depend on

time t. Robert Engle proposed the ARCH Model in 1982, and the generalized ARCH model, that is, the GARCH model, was proposed by Bollerslev in 1986 and various modified models such as EGARCH, GARCH-M, and GJR have been presented since then. The reason why many modified models of ARCH have been presented and received a lot of attention is because with the development of financial economics, not only time series prediction but also measurement and prediction of volatility (variance) have become very important. In many capital asset pricing models, the volatility of the underlying asset has been considered an important factor affecting the price of the asset, but the methodology for measuring and predicting this volatility had not been developed before ARCH. Research related to this has continued steadily not only in the 1980s, but also since then, and even recently, the GARCH model is now used in various engineering fields, especially network traffic analysis, in addition to financial economics.

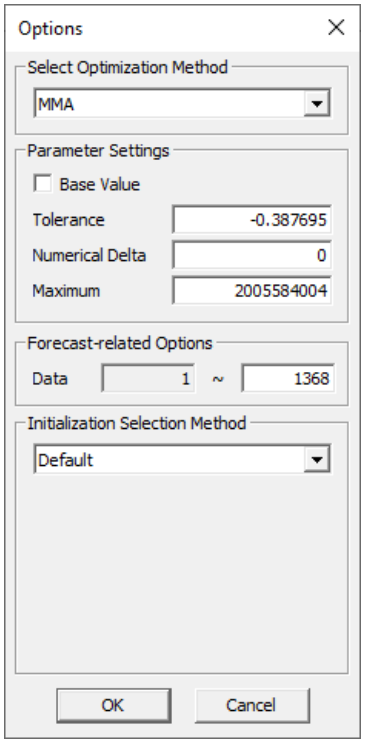
How to run

[Analyze] – [Time Series Analysis] – [Time Series Models] – [GARCH]

On the main screen, enter the basic information required for GARCH analysis. Select Dependent Variable (Response Variable) and Independent Variable (Explanatory Variable) and select ARCH order and GARCH order. Finally, choose whether you want to generate predictions or not and you're all set.



More specific settings can be set in the Option window below.



GARCH's Parameter Estimation is very demanding. In some cases, Estimation frequently ends at Local Optimum. Therefore, ECMiner™ presents several optimization methodologies for users to choose from. Users can perform optimization using multiple methodologies and choose the best solution. Parameter settings are options to set when performing the optimization algorithm. Forecast-related options determine which data to use for modeling. The initial value selection method determines how to select the initial value.

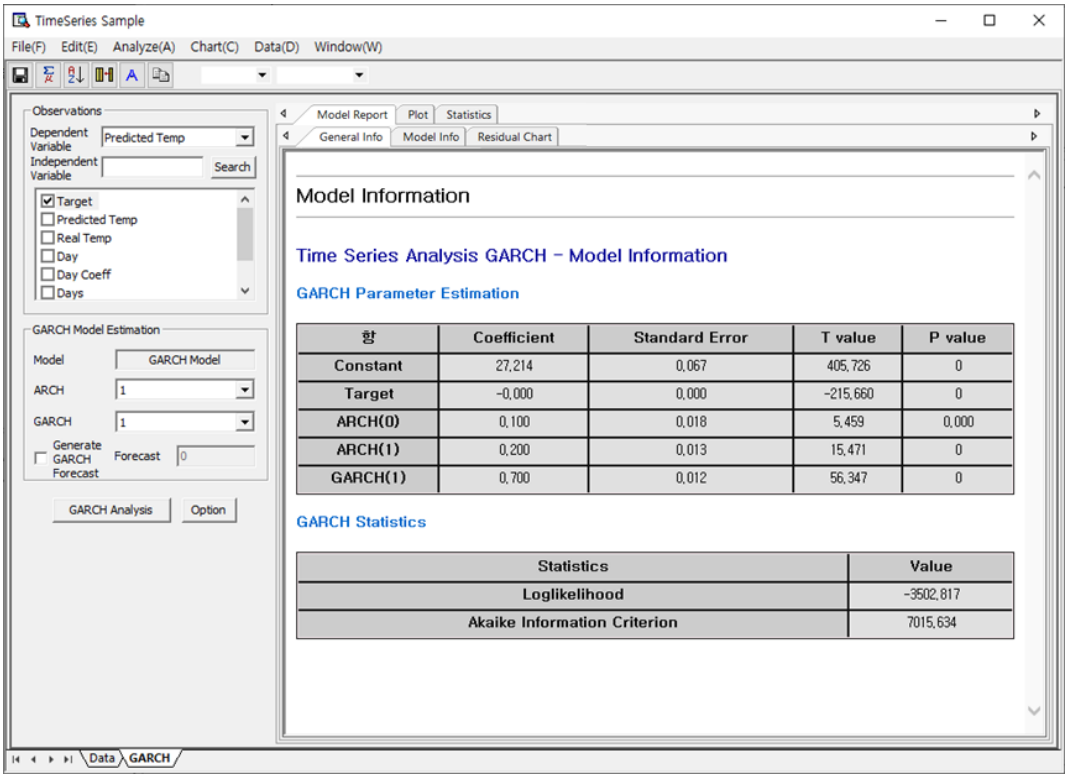
Results

- **Model Report**

  General Info: Shows basic information about time series data.

  Model Info: Provides parameters, statistics, and forecast results obtained through GARCH analysis.

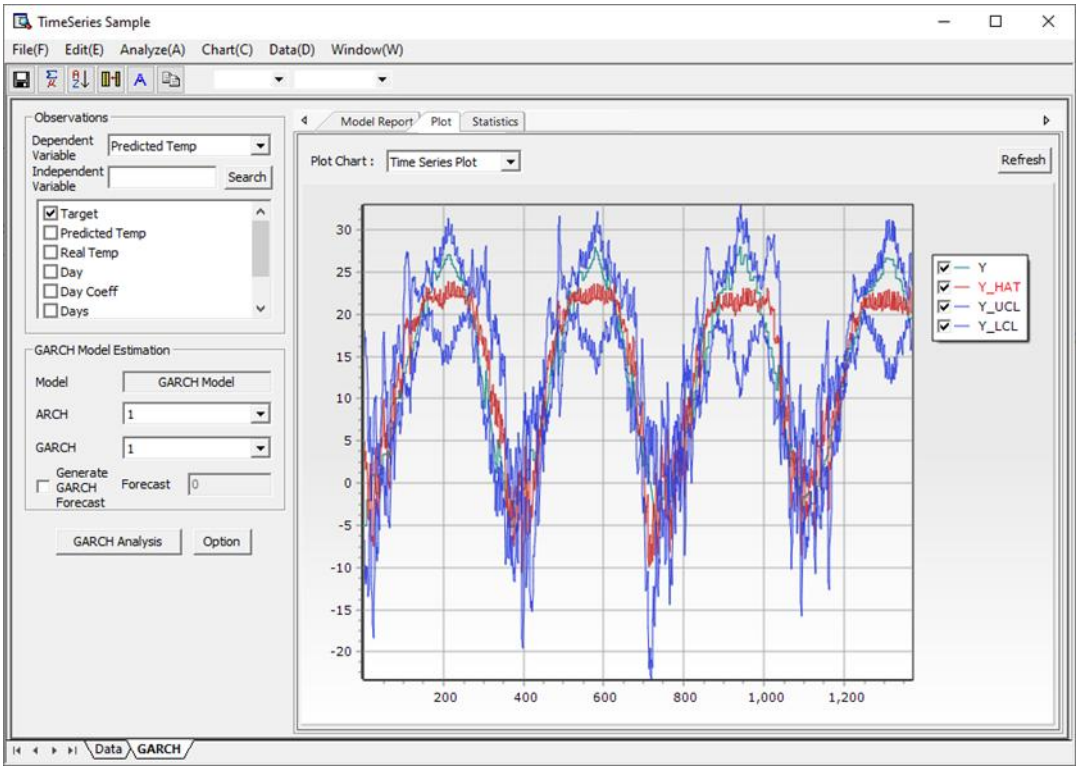
  Residual Chart: Shows Residual Autocorrelation Function, Residual Partial Autocorrelation Function.



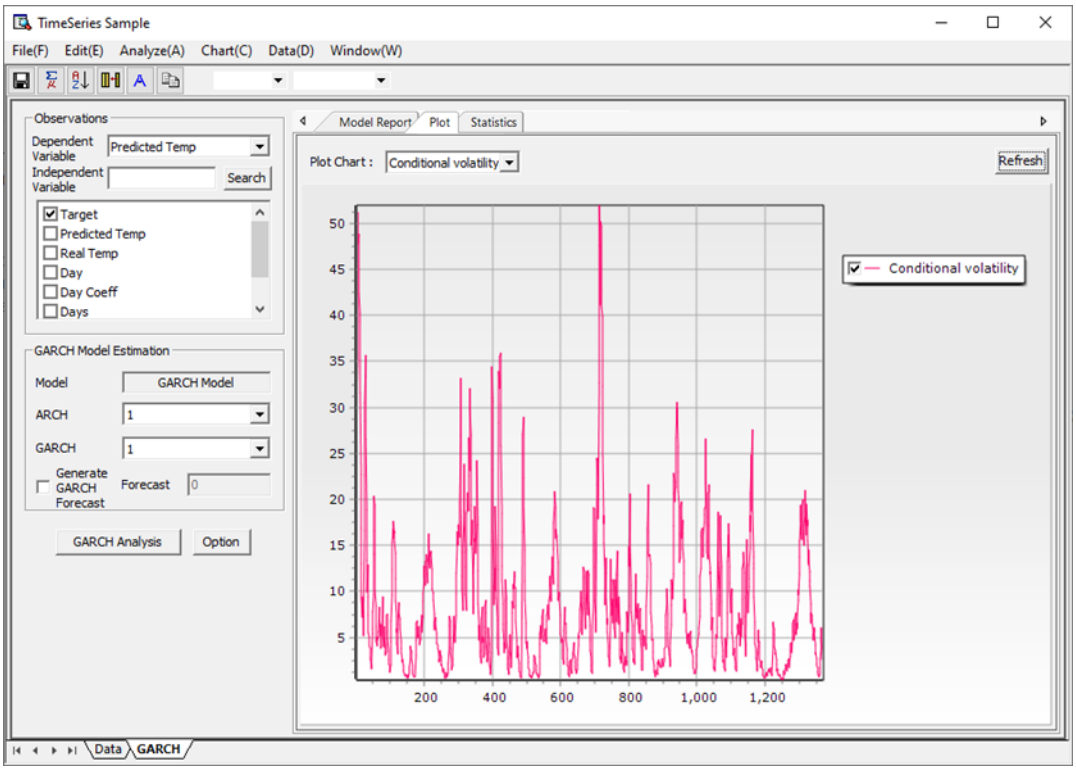
- **Plot**

  Visually displays data obtained through GARCH analysis. Shows Time Series Plot, squared residual, conditional volatility, and plots for residuals (Histogram, Normal Probability Plot, Residual vs. Order, Residual vs. Fitted Values).

  Below is an example of a Time Series Plot. You can see time series data, fitted values, and the upper and lower limits of the fitted values at a glance.

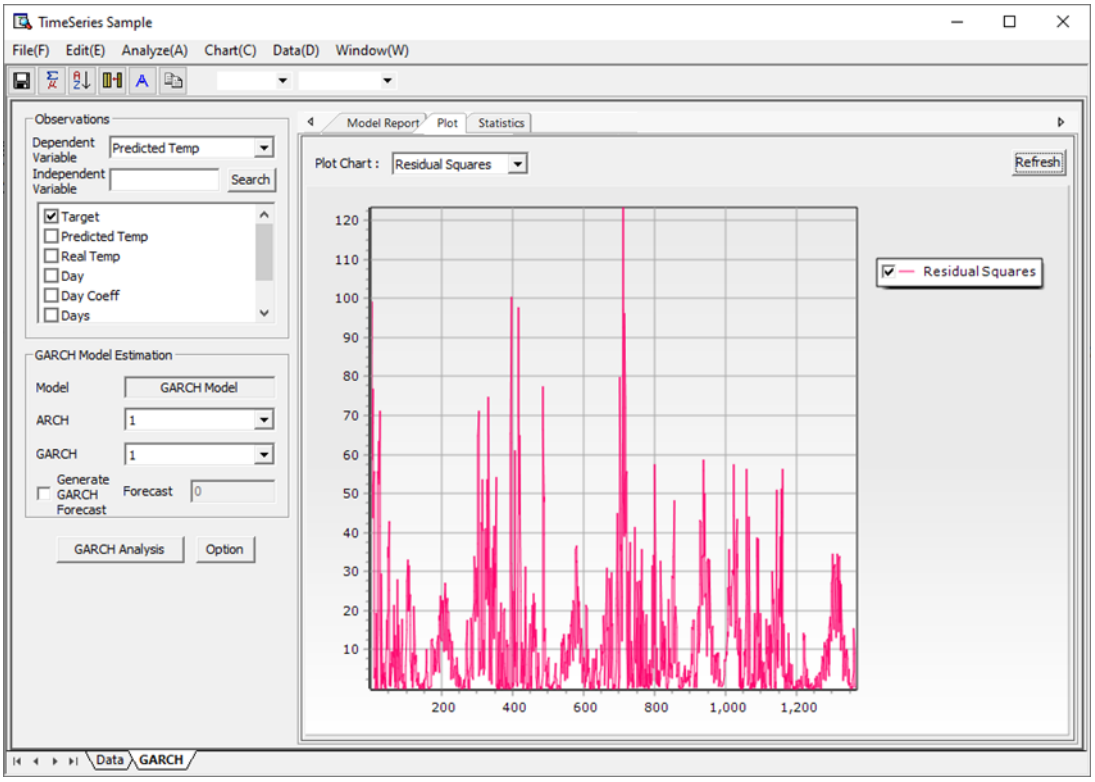


The plot below is a conditional volatility plot. You can view the conditional volatility value resulting from modeling. (If you make a prediction, the predicted value of conditional volatility will also appear.)

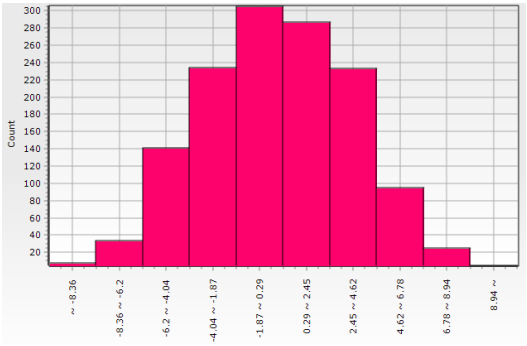
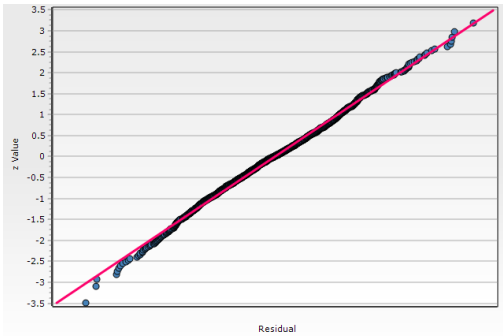


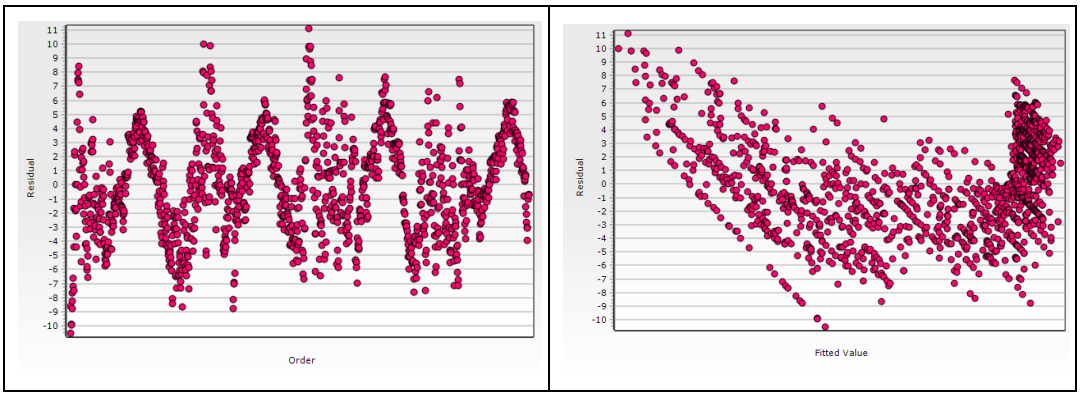
The plot below is a squared residual plot. You can see the value of the squared residual,

as a result of modeling.



In addition, you can check normality through four residual plots.

Histogram	Normal Probability Plot
	
Residual vs. Order	**Residual vs. Fitted Values**



- **Statistics**

  Statistics obtained through GARCH analysis can be seen in the table. In addition, it also provides the function to save tables.

---

(7) VAR

Overview

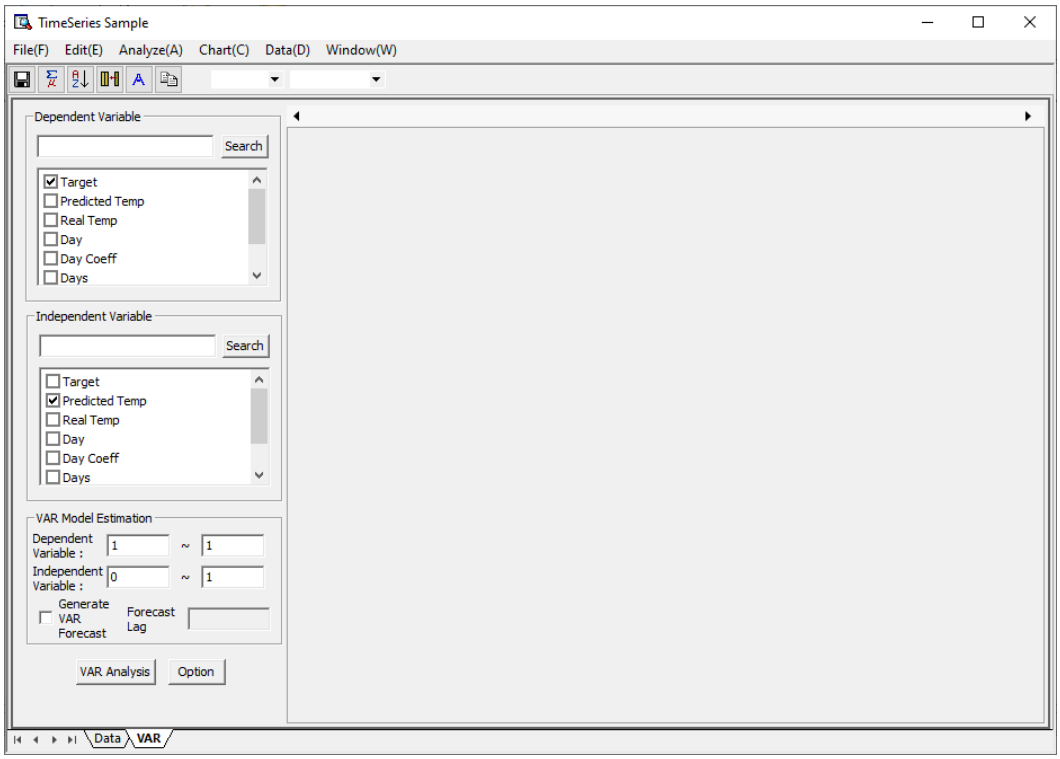
In empirical analysis, it is often advantageous to model two or more time series simultaneously. If sets of specific variables do not simply move individually but are influenced by each other, a model can be assumed as follows. This is called the Vector Autoregressive (VAR) model.

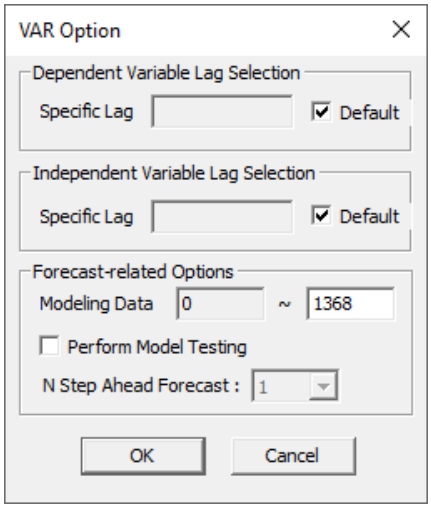
$$y_t = c + A_1 y_{t-1} + A_2 y_{t-2} + \cdots + A_p x_{t-p} + B_0 x_t + B_1 x_{t-1} + B_q x_{t-q} + \epsilon_t$$

How to run

[Analyze] – [Time Series Analysis] – [Time Series Models] – [VAR]

On the main screen, enter the basic information required for VAR analysis. When selecting Dependent Variable (Response Variable) and Independent Variable (Explanatory Variable), you can select multiple variables as dependent variables, and you may or may not select Independent Variable. When entering the degree of a dependent variable and the degree of an independent variable, enter the starting and ending degrees.





If you select default for **Dependent Variable** Lag Selection, the order set on the main screen is set. If you do not select default option, user can enter Specific Lag. (At this time, the order is divided by space.)

If you select default for **Independent Variable** Lag Selection, the order set on the Main screen is set. If you do not select default option, user can enter Specific Lag. (At this time, the order is divided by space.)

In the forecast-related options, you can select the data to be used for modeling and perform N step ahead forecast by selecting a specific step ahead when selecting to perform model verification. ECMiner™ performs N step ahead forecast starting immediately after the modeling data. Through this, analysts can assess whether the model created through modeling is useful for future predictions.

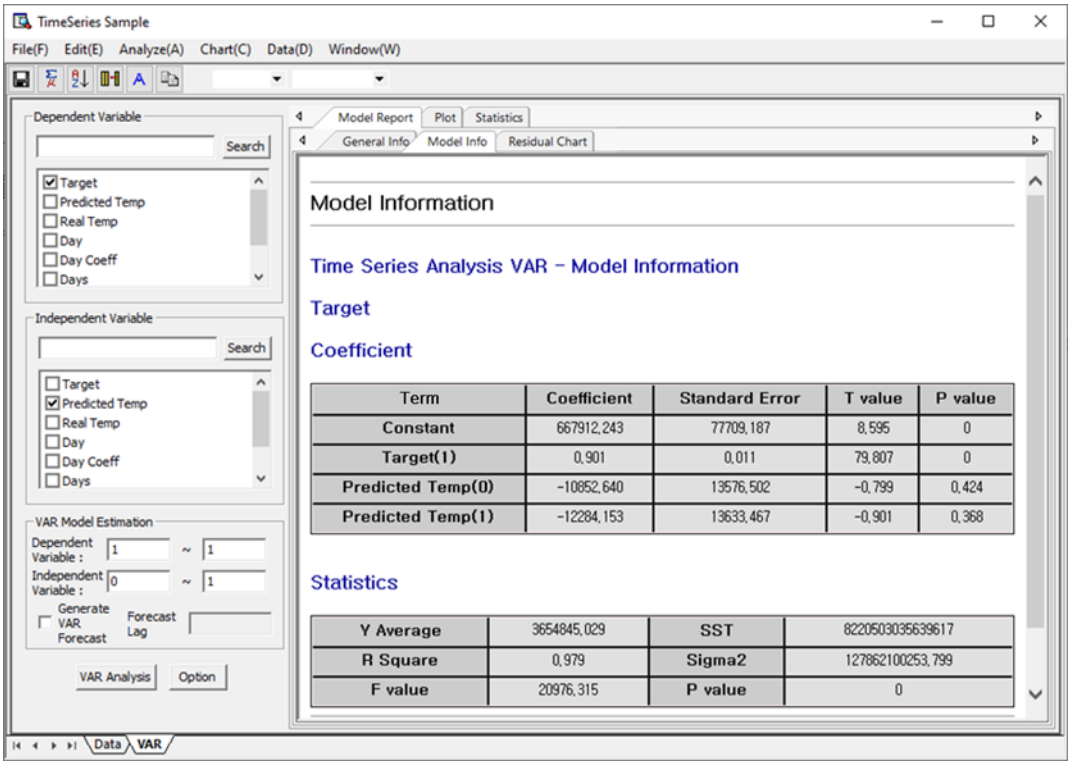
Results

- **Model Report**

  General Info: Shows basic information about time series data.

  Model Info: Provides parameters, statistics, and forecast results obtained through VAR analysis.

  Residual Chart: Shows Residual Autocorrelation Function, Residual Partial Autocorrelation Function.

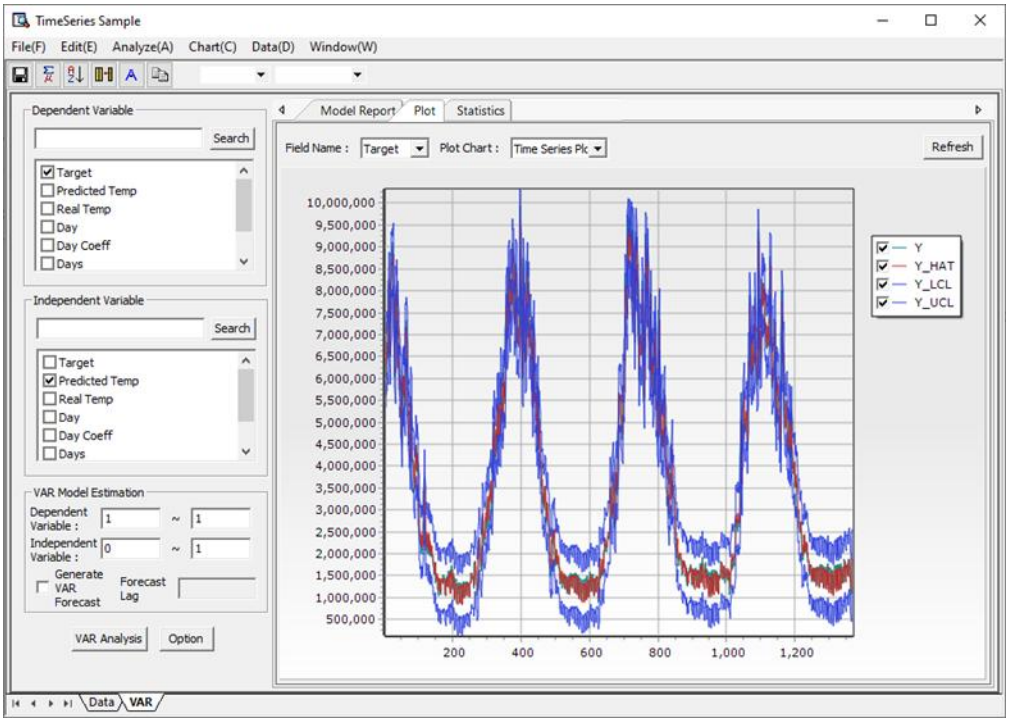


- **Plot**

  Visually displays data obtained through VAR analysis. Shows Time Series Plot and

Residual Plot (Histogram, Normal Probability Plot, Residual vs. Order, Residual vs. Fitted Values).

Below is an example of a Time Series Plot. You can see time series data, fitted values, and upper and lower limits of fitted values at a glance.



- **Statistics**

  The statistics obtained through VAR analysis can be seen in the table. In addition, it also provides the function to save tables.

---

(8) ARMAX

Overview

ARMAX stands for ARMA with Exogenous Variable and is a model that adds Exogenous Variable to the already mentioned ARMA model. This model can be described as:

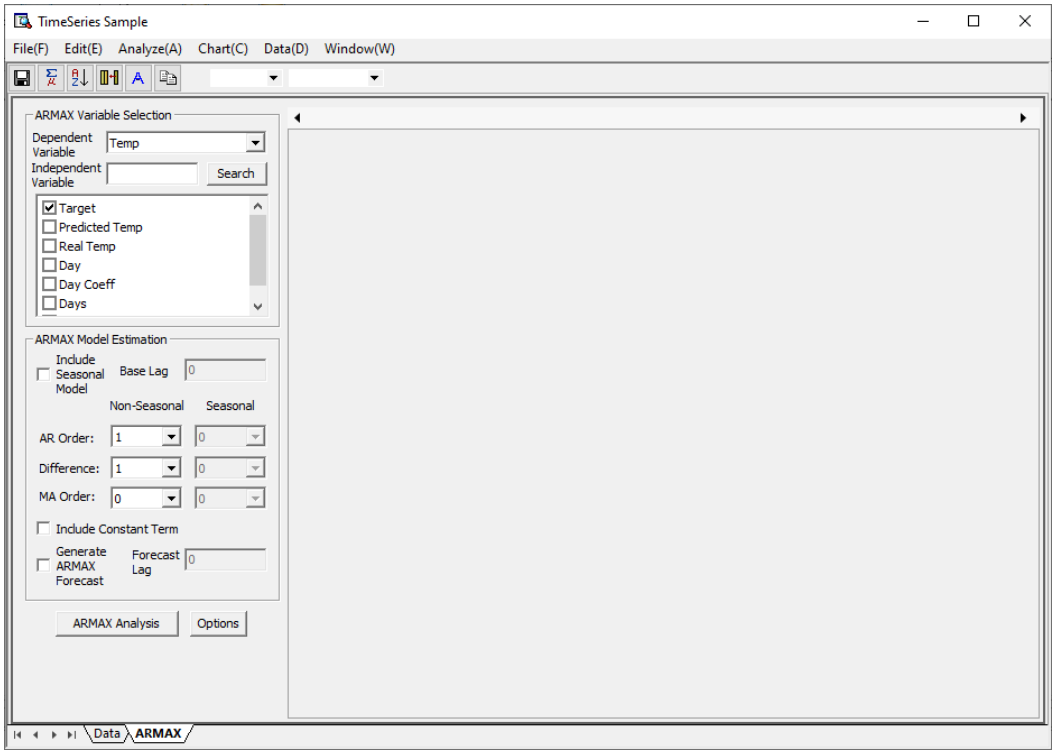
$$\phi(B)\left(y_t - \mu - x_{t1}\beta_1 - \cdots - x_{tm}\beta_m\right) = \theta(B)a_t$$

Both Parameter Estimation and Forecasting are the same as existing ARMA methods. All processes

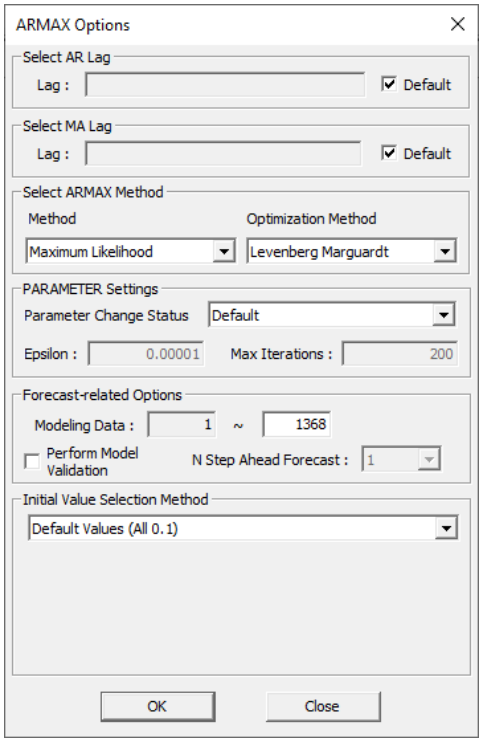
can be said to be the same if you just replace the existing y_t or $y_t - \mu$ with $y_t - x_{t1}\beta_1 - \cdots - x_{tm}\beta_m$ or $y_t - \mu - x_{t1}\beta_1 - \cdots - x_{tm}\beta_m$. This can be said to be a model that helps to better explain time series that cannot be explained simply with ARMA alone using Exogenous Variable.

How to run

[Analyze] – [Time Series Analysis] – [Time Series Models] – [ARMAX]



On main screen, you can enter basic information required for ARMAX analysis. When selecting Dependent Variable (Response Variable) and Independent Variable (Explanatory Variable), the independent variable may not be selected. In this case, you will get the same results as traditional ARIMA. Analysts can add seasonality or differences as in traditional ARIMA. ARMAX analysis can be performed through various order settings displayed on the main screen.



In addition, you can make more advanced settings through the ARMAX Options window. If you want to enter the AR order as the order specified by the analyst, uncheck the Default check box. Then, you can specify the AR order in space units. If you want to enter the MA order as the order specified by the analyst, uncheck the Default check box. Then you can specify the MA degree in space units.

As to which method to use to estimate the parameters of ARMAX, you can choose between Maximum Likelihood and Conditional Least Square. Levenberg Marquardt and Quasi Newton are provided as optimization methods. By using these two methods, users can use the better parameters of the two.

If you want to change the parameters used for optimization, set the Parameter Change Status as ‘Change’. This allows you to decide whether to optimize further or not.

Forecast-related Options allow you to select the data used for modeling and decide whether to perform model validation. When ‘Perform Model Validation’ is checked, the predicted values are calculated from the modeling data. Through this, you can gauge how accurate the model obtained through modeling is.

The Initial Value Selection Method is intended to compensate for the limitations of Nonlinear

Optimization. It helps to find the optimal solution by setting various initial values.

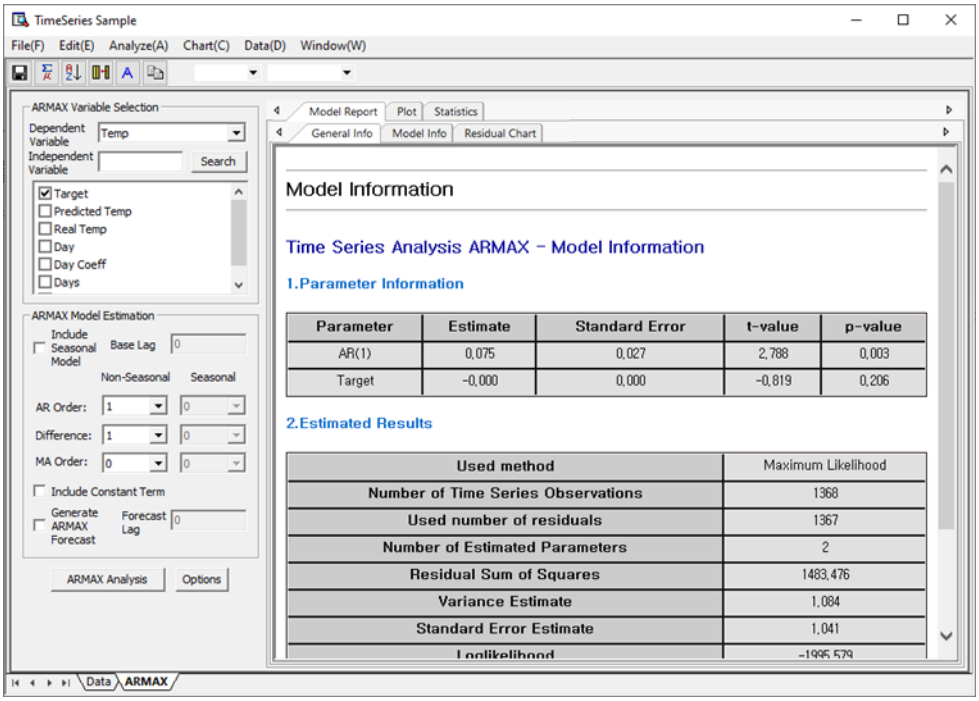
Results

- **Model Report**

  General Info: Shows basic information about time series data.

  Model Info: Provides parameters, statistics, and forecast results obtained through ARMAX analysis.

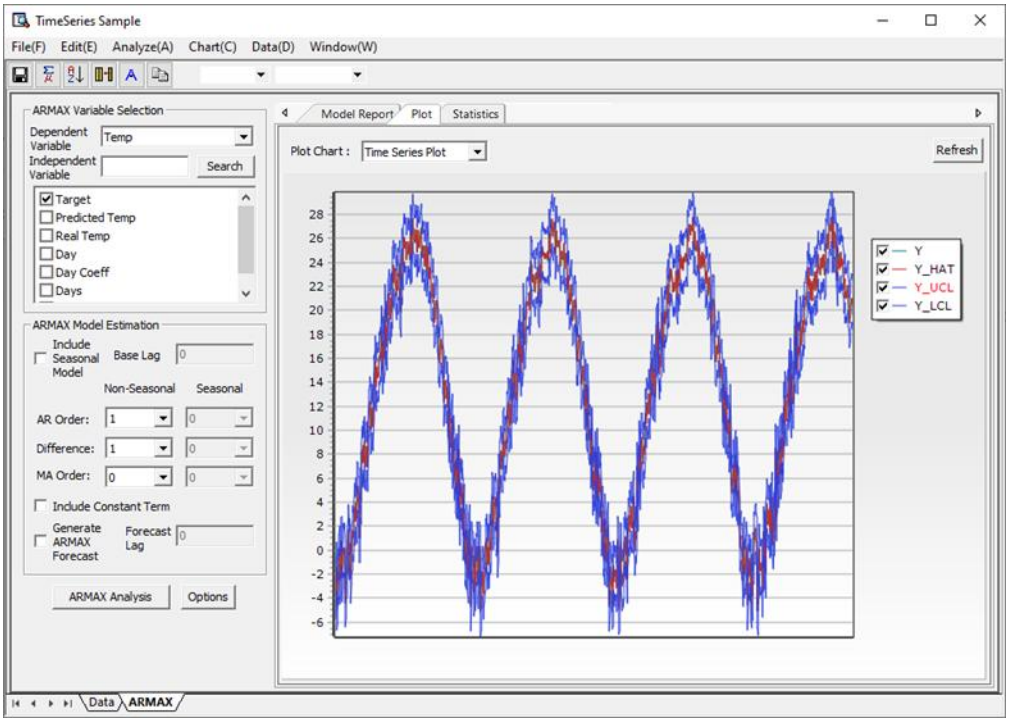
  Residual Chart: Shows Residual Autocorrelation Function, Residual Partial Autocorrelation Function.



- **Plot**

  Visually displays data obtained through ARMAX analysis. Shows Time Series Plot and Residual Plot (Histogram, Normal Probability Plot, Residual vs. Order, Residual vs. Fitted Values).

  Below is an example of a Time Series Plot. You can see time series data, fitted values, and upper and lower limits of fitted values at a glance.



- **Statistics**

  The statistics obtained through VAR analysis can be seen in the table. In addition, it also provides the function to save tables.

4.3.6.2 Time Series Test

---

(1) Unit Root Test

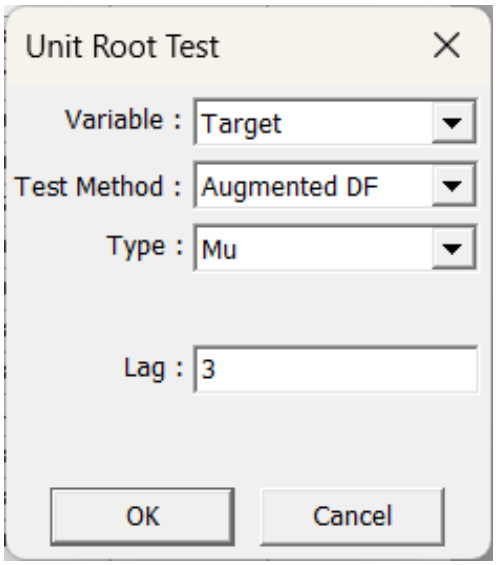
Overview

Non-stationary time series have a characteristic root of 1 when expressed as an autoregressive model. That is, it has a unit root. The Time Series Test menu supports KPSS (Kwiatkowski & Phillips & Schmidt & Shin) test and ADF (Augmented Dickey-Fuller) test. The null hypothesis of the ADF test is that a unit root exists in the variable, and the null hypothesis of the KPSS test is that the unit root does not exist in the variable. Decide whether to accept or reject the null hypothesis through test statistics (TAU) and significance probability (p-value).

How to use

[Analyze] – [Time Series Analysis] – [Time Series Test] – [Unit Root Test]

In the window, select which test method to use, select/enter the parameters required for each method, and then select the variable to be tested. Click the OK button and the Time Series Test will be performed.



- **ADF (Augmented Dickey-Fuller) Test**

  Time series data can be represented in different forms depending on whether constants and deterministic trends are included. The selection type of ADF Test must be entered according to the type of time series data to be tested.

  The test model for the case that does not include constants and deterministic trends (Type = none) is as follows.

$$y_t = \gamma y_{t-1} + \sum_{i=1}^{p} \nabla y_{t-i} + e_t$$

  The test model when only constants are included (Type = Drift) can be expressed as follows.

$$y_t = a + \gamma y_{t-1} + \sum_{i=1}^{p} \nabla y_{t-i} + e_t$$

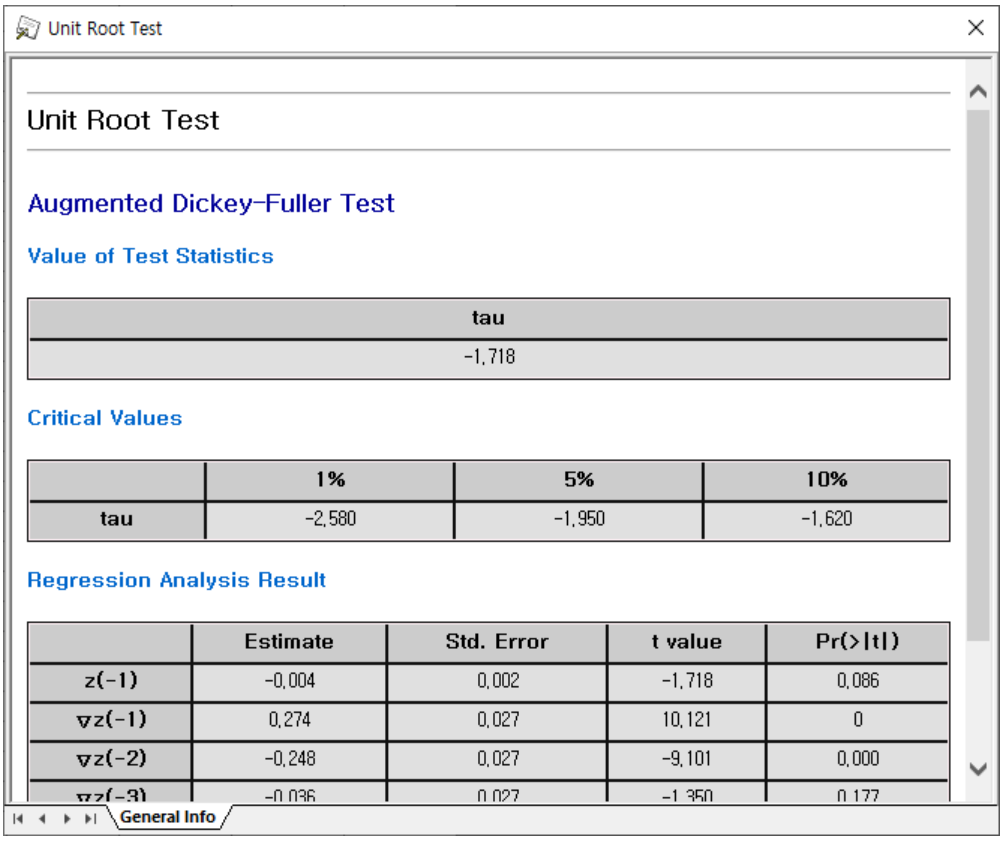
  The test model when including both constant and deterministic trends (Type = Trend) is

as follows.

$$y_t = a + \beta t + \gamma y_{t-1} + \sum_{i=1}^{p} \nabla y_{t-i} + e_t$$

Lag is the order of the first autoregressive term included in the regression equation of the test model, and is the p value of the above regression equation. So you can get the correct result, only when the Lag value is greater than 1.

Results



It provides statistics such as Test-statistics, Critical value, Coefficient, Multiple R-square, Adjusted R-square, F-statistics, p-value, etc. according to each test method. In Coefficient, Intercept is a constant term, tt is a deterministic trend, z.lag.1 is y, and z.diff.lag1 is the value of the first difference of y.

- **KPSS(Kwiatkowski & Phillips & Schmidt & Shin) Test**

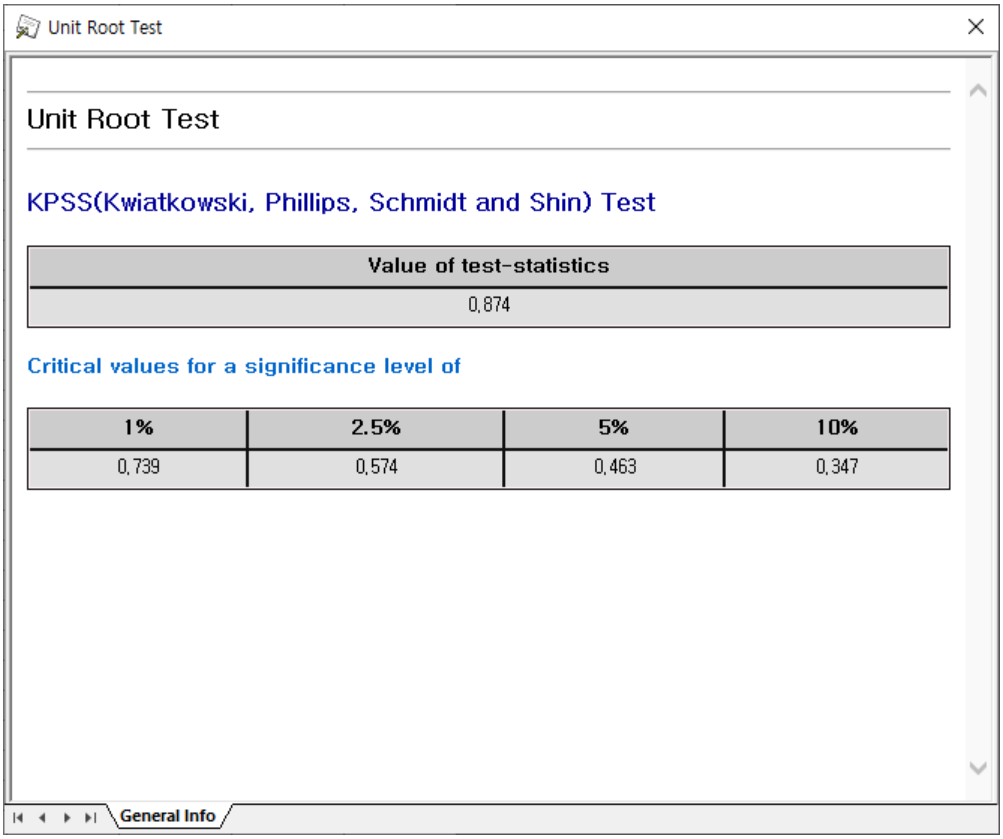
  The formula for calculating the test statistics value of KPSS is as follows.

$$KPSS = N^{-2}\sum_{t=1}^{N} S_t^2 \Big/ \hat{\sigma}^2(p)$$

When a time series is expressed as a regression equation of random work + deterministic trend + stationary error like $X_t = r_t + \beta t = \varepsilon_t$, then $S_i = \sum_{j=1}^{t} e_j$. (et , t = 1, 2, … , N, are the residuals)

If Type is entered as Mu, set residuals as 'residual = y – mean(y)', and if Type is entered as Tau, uses residuals from the result of linear regression which has y as dependent variable, and time trend as independent variable.

Results



This test method provides test statistics and critical value values for each significance level (1%, 2.5%, 5%, 10%).

---

(2) Granger Causality

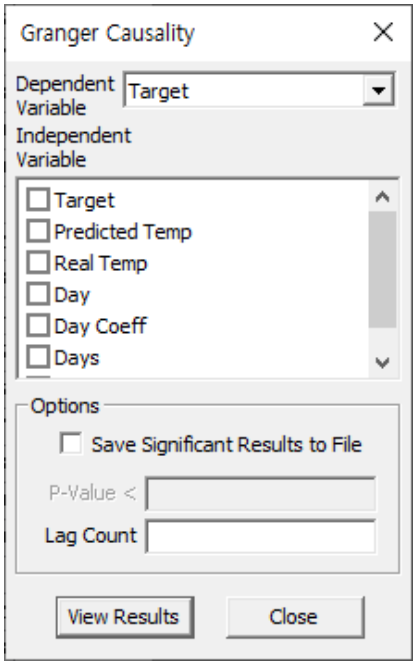
Overview

The Granger causality test is a statistical hypothesis test that tests whether the one time series is useful for predicting the other time series. Through t-test and F-test, we test whether Independent Variable provides statistically significant information about the future of Dependent Variable.

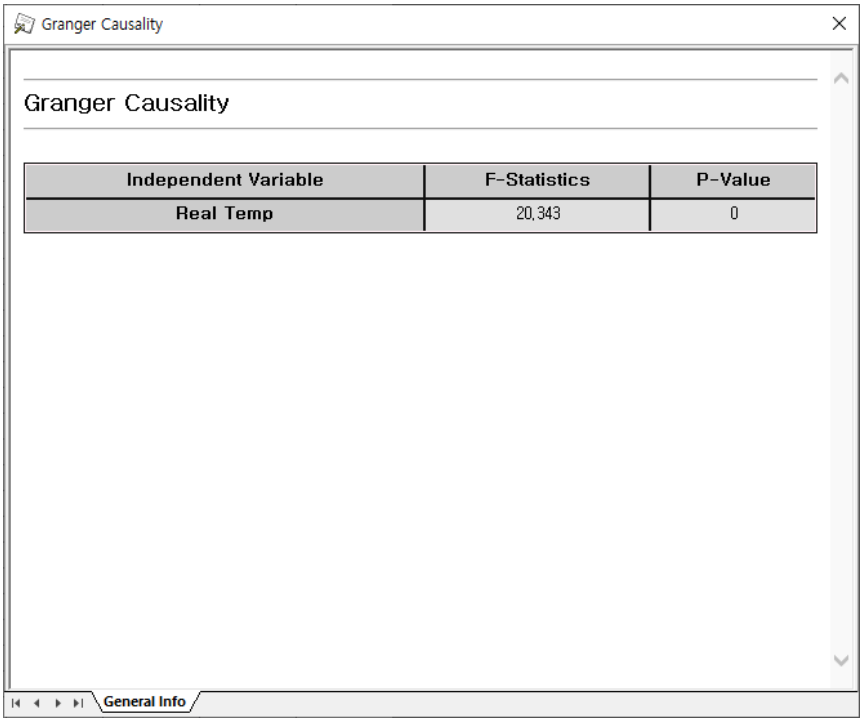
How to run

[Analyze] – [Time Series Analysis] – [Time Series Test] – [Granger Causality]

In the window select Dependent Variable and Independent Variable. If you check “Save Significant Results to File”, you can save the test result, a list of independent variables whose p-value is less than the value entered by the user, and the p-value value as a text file. Click the View Results button to perform Granger causality.



Results



Provides F-Statistics value and p-Value, which are causality test results.

---

(3) Cointegration Test

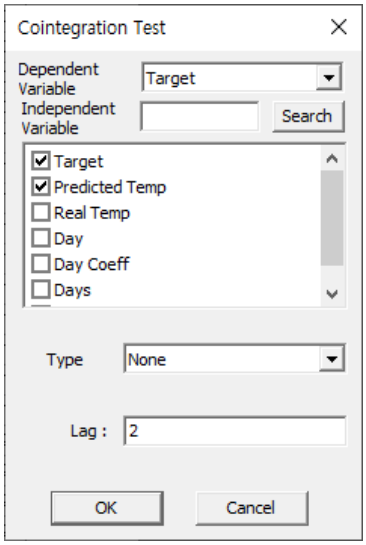
Overview

When two or more time series are individually integrated, but some of their linear combinations are low order, these time series are said to be cointegrated. Cointegration Test tests how cointegrated two or more time series are. There are methods such as The Engle-Granger two-step method, The Johansen's procedure, and Phillips-Ouliaris Cointegration Test, and the Johansen's procedure is used in the Cointegration Test menu.

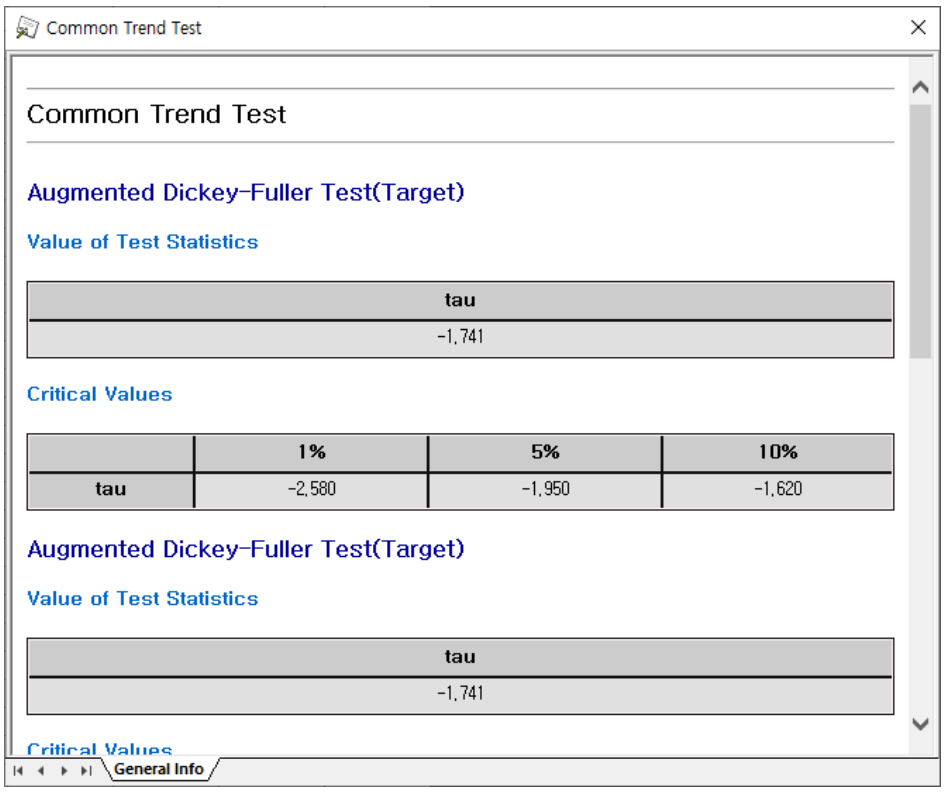
How to run

[Analyze] – [Time Series Analysis] – [Time Series Test] – [Cointegration Test]

Select the dependent and independent variables in the window. In the Type menu, you can select None, Constant, or Trend.



Results



Provides the value of test statistics, which is the result of the cointegration test.

---

(4) ARCH Test

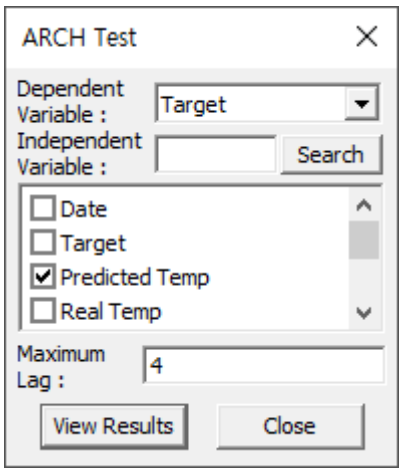
Overview

The ARCH model is a heteroskedastic conditional autoregressive model for predicting the volatility of a time series that changes over time. The ARCH test is used to determine whether the current variance can be predicted by the past variance.

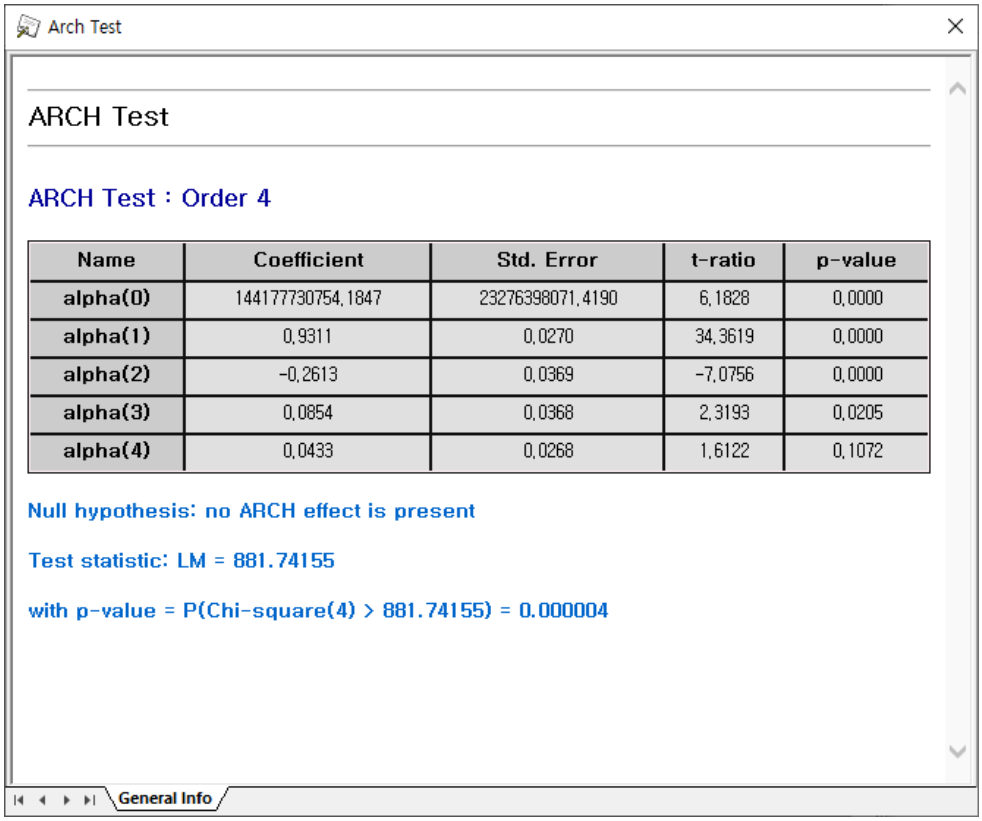
How to run

[Analyze] – [Time Series Analysis] – [Time Series Test] – [ARCH Test]

Select the dependent and independent variables in the window. Enter the maximum disparity value. Arch Test is performed by clicking the View Results button.



Results



Test statistics of Arch test are provided.

4.3.6.3 Time Series Correlation

---

(1) Cross-Correlation

Overview

Cross-Correlation is an indicator that measures the similarity between two time series. Cross-Correlation is mainly used to discover short, known features in long-duration signals. When there

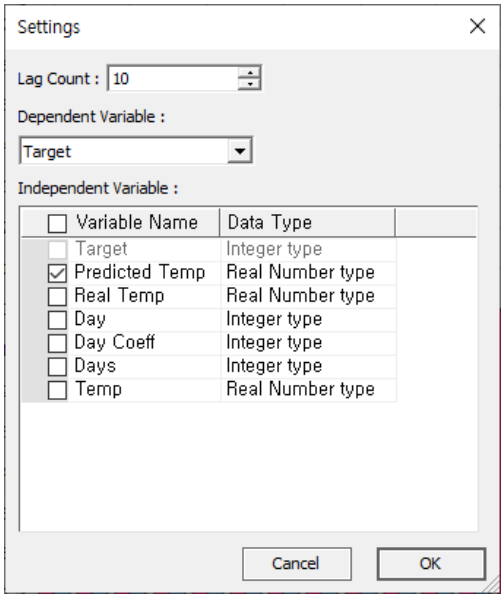
are two time series X and Y, the k-time difference cross-correlation coefficient r_k between the two-time series is calculated as follows.

$$r_k = \frac{\sum(X_t - \bar{X})\,(Y_{t+k} - \bar{Y})}{\sqrt{\sum(X_t - \bar{X})^2 \sum(Y_t - \bar{Y})^2}}$$

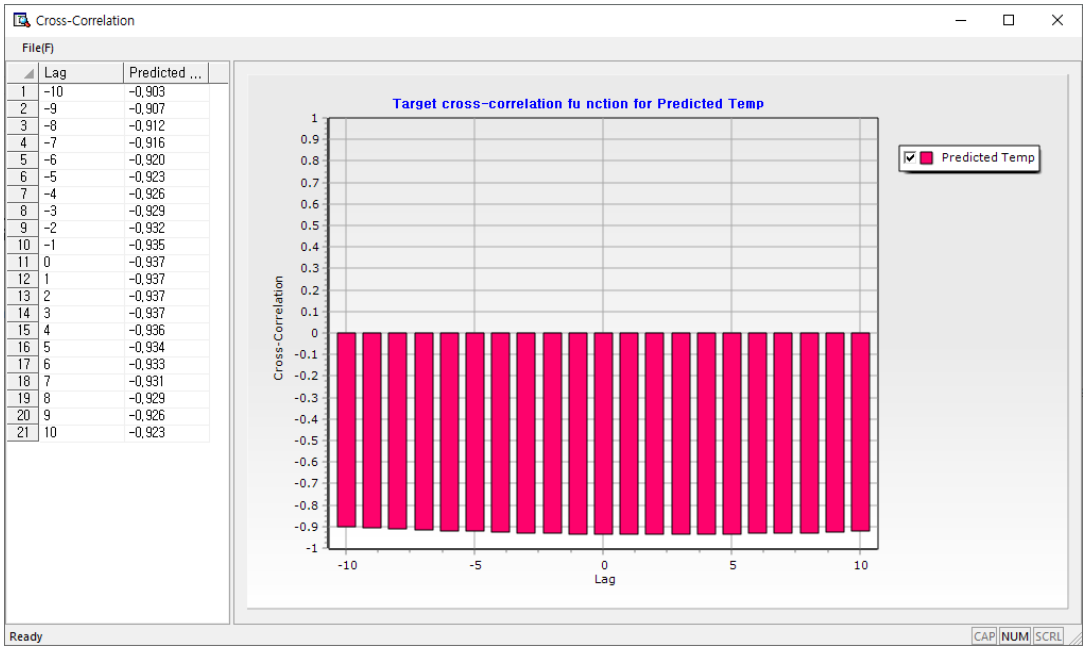
How to run

[Analyze] – [Time Series Analysis] – [Time Series Correlation] – [Cross-Correlation]

Open the Settings dialog from the menu “File>Settings”. Select dependent and independent variables in the window. Cross-Correlation is executed when you click the OK button.



Results



Provides the Cross-Correlation coefficient between Independent Variable and Dependent Variable

at the corresponding time difference.

---

(2) Autocorrelation, Partial Autocorrelation

Overview

Autocorrelation is a measure that shows how the value of current time series data is correlated with past data. The higher autocorrelation values mean the higher the possibility that predict your current value through your past data.

To explain the partial autocorrelation, we will use the following model as an example.

$$z_t = \phi_1 z_{t-1} + a_t \quad a_t \sim iid\ N(0, \sigma^2)$$

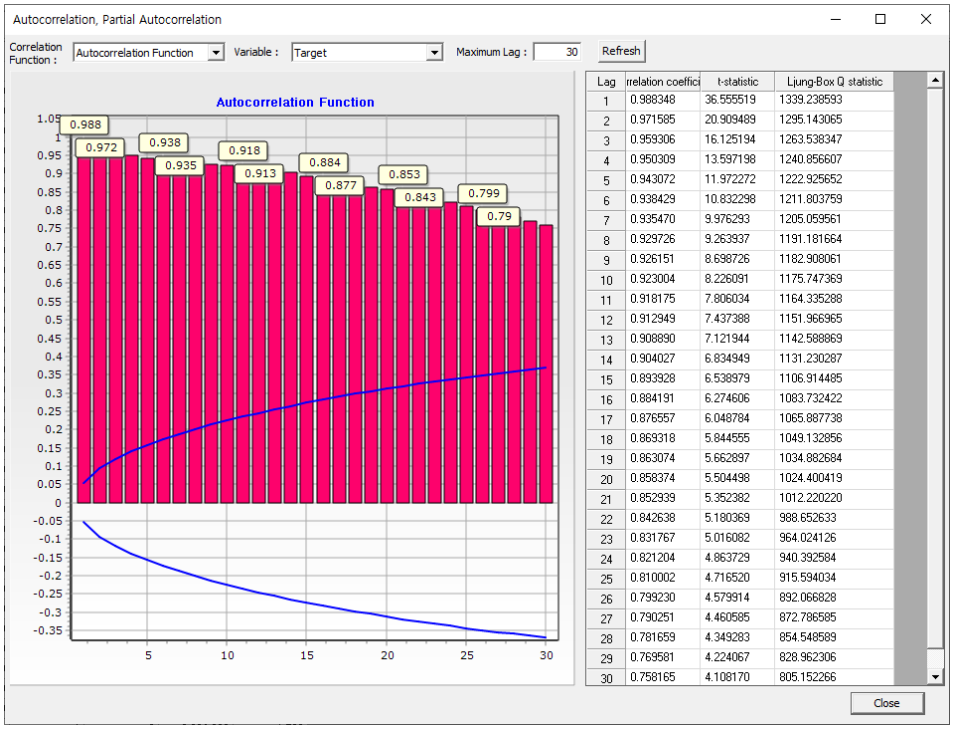
If you calculate the second-order autocorrelation coefficient from the above time series model, you will get a significant value and the analyst may think that the autocorrelation coefficient in the second order is significant because the above time series was created in AR(2), or because it was created in AR(3). Therefore, in addition to autocorrelation, the concept of partial autocorrelation is needed. For example, if you want to find the second-order partial autocorrelation coefficient, you find the correlation coefficient after excluding the influence of the first-order. In this case, the value of the second-order autocorrelation coefficient may be high, but the value of the second-order partial autocorrelation function will be small.

How to run

[Analyze] – [Time Series Analysis] – [Time Series Correlation] – [Autocorrelation, Partial Autocorrelation]

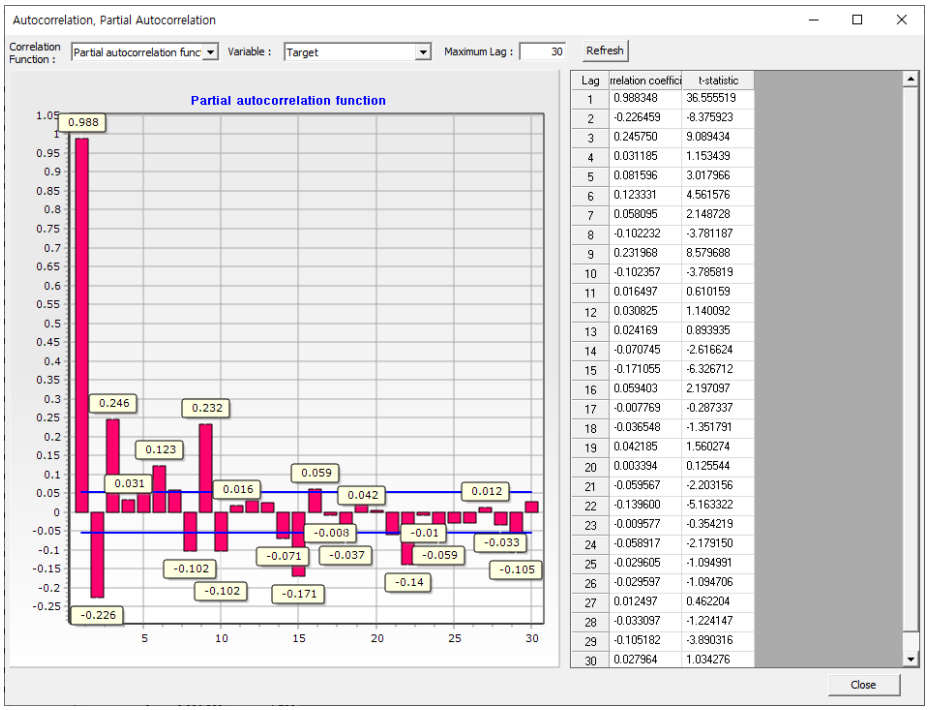
Select the type of correlation function in the window. The analyst can select the autocorrelation function or partial autocorrelation function and select the variable to be analyzed along with it. By determining the maximum number of lags, the analyst can determine how much lag the correlation function will be obtained.

Results



Lag	rrelation coeffici	t-statistic	Ljung-Box Q statistic
1	0.988348	36.555519	1339.238593
2	0.971585	20.909489	1295.143065
3	0.959306	16.125194	1263.538347
4	0.950309	13.597198	1240.856607
5	0.943072	11.972272	1222.925652
6	0.938429	10.832298	1211.803759
7	0.935470	9.976293	1205.059561
8	0.929726	9.263937	1191.181664
9	0.926151	8.698726	1182.908061
10	0.923004	8.226091	1175.747369
11	0.918175	7.806034	1164.335288
12	0.912949	7.437388	1151.966965
13	0.908890	7.121944	1142.588869
14	0.904027	6.834949	1131.230287
15	0.893928	6.538979	1106.914485
16	0.884191	6.274606	1083.732422
17	0.876557	6.048784	1065.887738
18	0.869318	5.844555	1049.132856
19	0.863074	5.662897	1034.882684
20	0.858374	5.504498	1024.400419
21	0.852939	5.352382	1012.220220
22	0.842638	5.180369	988.652633
23	0.831767	5.016082	964.024126
24	0.821204	4.863729	940.392584
25	0.810002	4.716520	915.594034
26	0.799230	4.579914	892.066828
27	0.790251	4.460585	872.786585
28	0.781659	4.349283	854.548589
29	0.769581	4.224067	828.962306
30	0.758165	4.108170	805.152266

Above is an example of an autocorrelation function. The autocorrelation function shows the correlation coefficient and t-statistic as well as the Ljung-Box Q-test statistic. The Q test statistic value is used to test the hypothesis of whether data are independently distributed. Independently distributed data means that any correlation found in the data was created by random sampling.



Lag	rrelation coeffici	t-statistic
1	0.988348	36.555519
2	-0.226459	-8.375923
3	0.245750	9.089434
4	0.031185	1.153439
5	0.081596	3.017966
6	0.123331	4.561576
7	0.058095	2.148728
8	-0.102232	-3.781187
9	0.231968	8.579688
10	-0.102357	-3.785819
11	0.016497	0.610159
12	0.030825	1.140092
13	0.024169	0.893935
14	-0.070745	-2.616624
15	-0.171055	-6.326712
16	0.059403	2.197097
17	-0.007769	-0.287337
18	-0.036548	-1.351791
19	0.042185	1.560274
20	0.003394	0.125544
21	-0.059567	-2.203156
22	-0.139600	-5.163322
23	-0.009577	-0.354219
24	-0.058917	-2.179150
25	-0.029605	-1.094991
26	-0.029597	-1.094706
27	0.012497	0.462204
28	-0.033097	-1.224147
29	-0.105182	-3.890316
30	0.027964	1.034276

Above is an example of partial autocorrelation function.

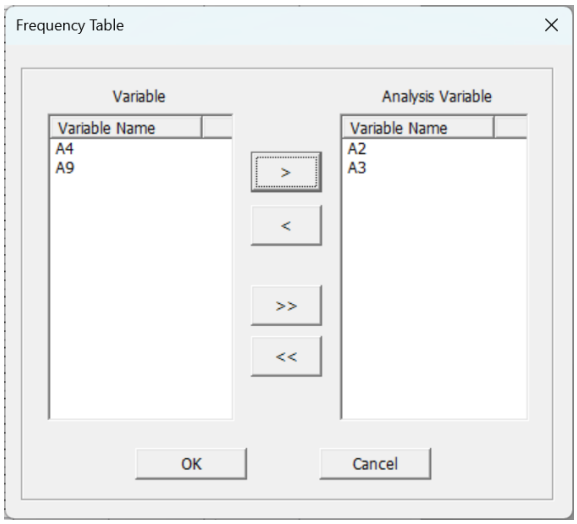
4.3.7 Table

4.3.7.1 Frequency Table

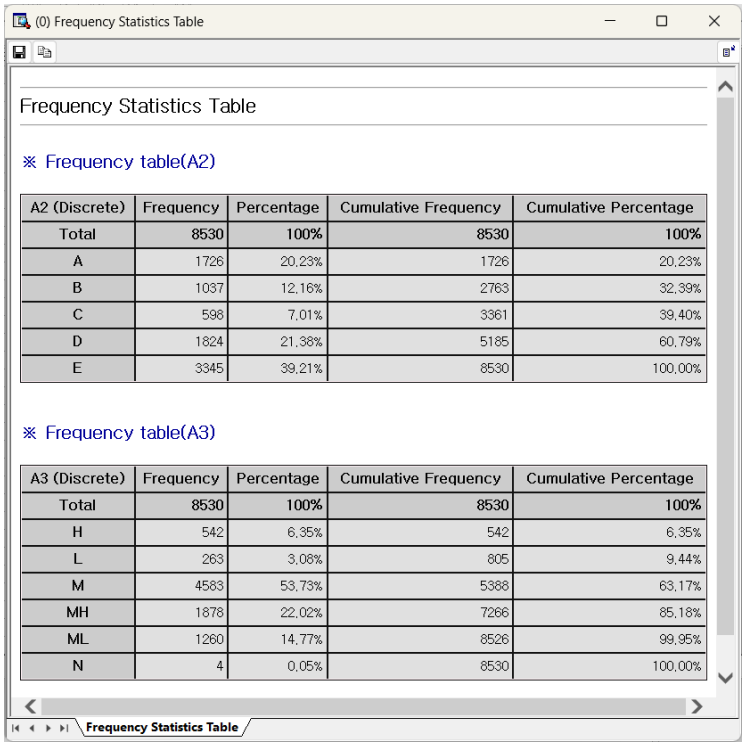
Frequency Table shows how often each distinct value occurs in a dataset.

How to run

[Analyze] – [Table] – [Frequency Table]



Results



A2 (Discrete)	Frequency	Percentage	Cumulative Frequency	Cumulative Percentage
Total	8530	100%	8530	100%
A	1726	20,23%	1726	20,23%
B	1037	12,16%	2763	32,39%
C	598	7,01%	3361	39,40%
D	1824	21,38%	5185	60,79%
E	3345	39,21%	8530	100,00%

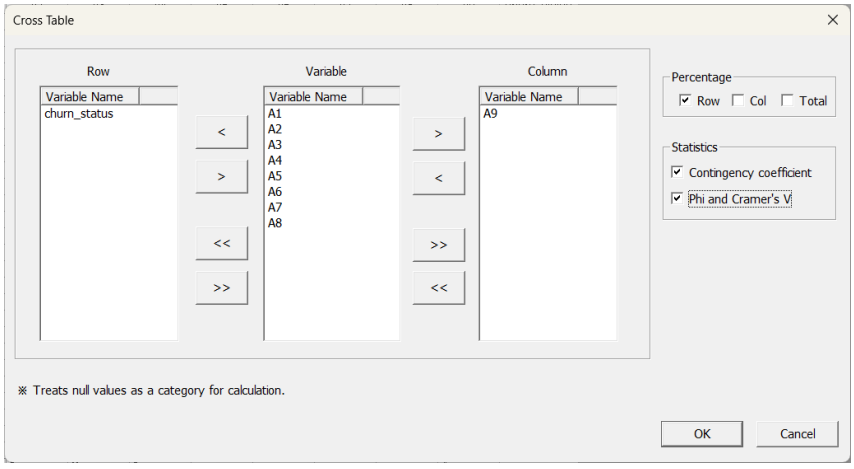
A3 (Discrete)	Frequency	Percentage	Cumulative Frequency	Cumulative Percentage
Total	8530	100%	8530	100%
H	542	6,35%	542	6,35%
L	263	3,08%	805	9,44%
M	4583	53,73%	5388	63,17%
MH	1878	22,02%	7266	85,18%
ML	1260	14,77%	8526	99,95%
N	4	0,05%	8530	100,00%

4.3.7.2 Cross Table

Cross table is for two variables.

How to run

[Analyze] – [Table] – [Cross Table]



Percentage options are row percentage, column percentage and total percentage.

Statistics options measure the strength of the association between row and column variables based on chi-square test, and provide Contingency coefficient and Phi and Cramer's V. Both statistics have values between 0 and 1, with the closer to 1 indicating a stronger association.

Results

Cross-Tabulation

※ Cross-tabulation (churn_status vs. A9)

		A	B	C	D	Total
0	Count % in churn_status	3780 98.28%	35 0.91%	7 0.18%	24 0.62%	3846 100%
1	Count % in churn_status	4680 99.91%	2 0.04%	0 0.00%	2 0.04%	4684 100%
Total	Count % in churn_status	8460 99.18%	37 0.43%	7 0.08%	26 0.30%	8530 100%

※ Nominal statistics

		Value
Nominal - Nominal	Phi	0.09003
	Cramer's V	0.09003
	Contingency Coefficient	0.08966
Total data count		8530

* When the category count is not the same, the Cramers'V statistic is most appropriate as nominal statistics.

4.3.7.3 Univariate Chi-Square Test

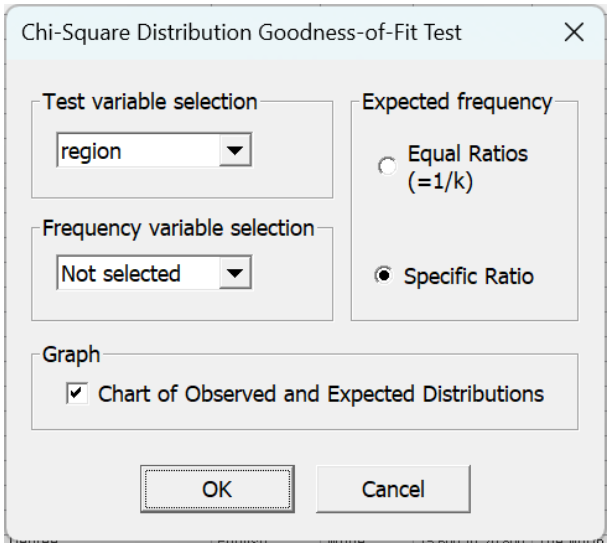
Univariate Chi-Square Test checks if the observed data fits a specified distribution with expected frequencies or ratios.

How to run

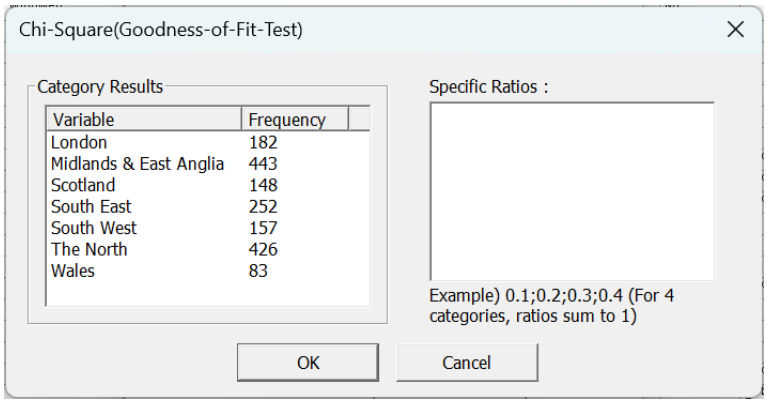
[Analyze] – [Table] – [Univariate Chi-Square Test]

Select the Test variable (categorical variable) and frequency variable, and select Equal Ratios or

Specific Ratio. You can select a chart of observed and expected distribution.



Specific Ratio refers to the frequency by category, where you enter the ratios directly. You must enter a ratio for each category, separated by ';'.



Results

The results of the chi-square goodness-of-fit test appear as follows:

The contribution to the test ratio, expected value, and chi-square test statistic for each category is displayed, and the p-value is used to test the data's fit to the expected frequency.

Chi-Square(Goodness-of-Fit-Test)

Chi-Square Goodness-of-Fit Test Variable: region

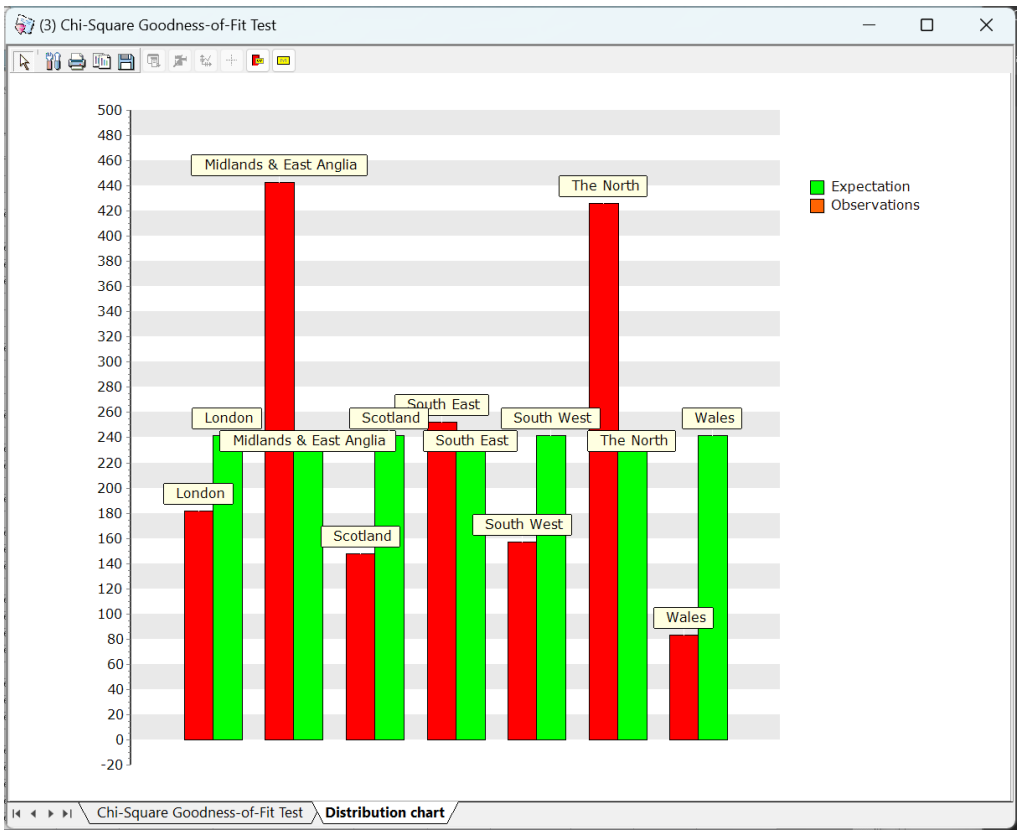
N = 1691

Category	Number of Observations	Test Ratio	Expected Value	Contribution to χ2
London	182	0.14286	241.57143	14.69029
Midlands & East Anglia	443	0.14286	241.57143	167.95641
Scotland	148	0.14286	241.57143	36.24440
South East	252	0.14286	241.57143	0.45020
South West	157	0.14286	241.57143	29.60750
The North	426	0.14286	241.57143	140.80265
Wales	83	0.14286	241.57143	104.08887

Test Statistics

χ2	DF	p-value
493.84033	6	0.00000

Select the distribution chart tab in the results window to view the observed and expected values distribution chart



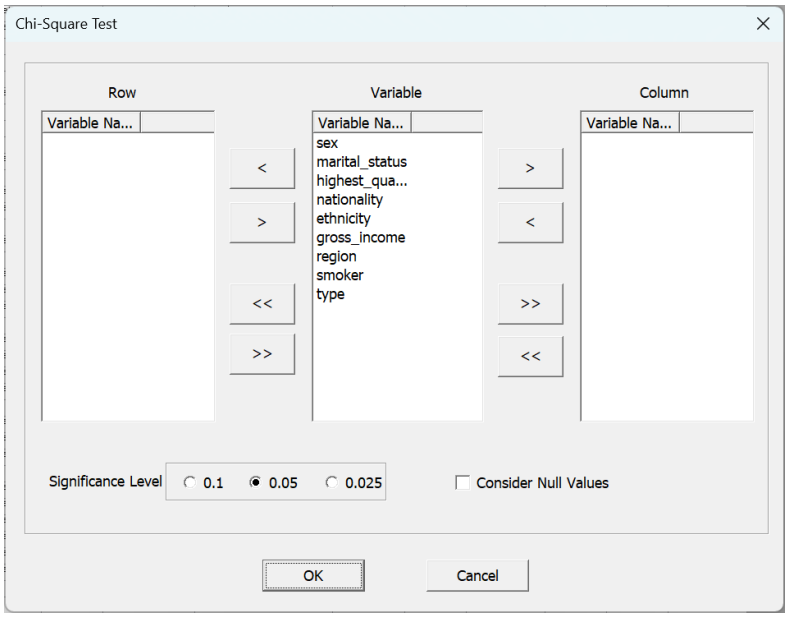
4.3.7.4 Independence Test (Chi-Square Test)

The Independence Test – Chi-Square Test is to test whether two categorical variables are independent.

How to run

[Analyze] – [Table] – [Independence Test]

Select two variables to analyze



Results

Chi-Square Test

※ Cross-tabulation (sex vs. smoker)

		No	Yes	Total
Female	Count % in sex % in smoker % in Total	731 75.75% 57.56% 43.23%	234 24.25% 55.58% 13.84%	965 100% 57.07%
Male	Count % in sex % in smoker % in Total	539 74.24% 42.44% 31.87%	187 25.76% 44.42% 11.06%	726 100% 42.93%
Total	Count % in sex % in smoker % in Total	1270 100% 75.10%	421 100% 24.90%	1691 100%

※ Statistics (sex vs. smoker)

Statistics	Value	DF	p-value
Pearson Chi-Square	0.50446	1	0.47755
Likelihood Chi-Square	0.50356	1	0.47794

4.3.8 Probability Distribution

4.3.8.1 Parameter Estimation

There are several methods to estimate the parameters of a probability distribution, including Method of Moment and Maximum Likelihood. ECMiner™ optimizes and estimates the parameters of a probability distribution using the Maximum Likelihood method based on the given data, and it provides confidence intervals for the estimated parameters to indicate their reliability.

How to run

[Analyze] – [Probability Distribution] – [Parameter Estimation]

There are 12 distributions. The user can select the distribution to fit the data, specify which field contains the data, and choose the desired confidence level for the estimation. After making these selections, clicking "View Results" will display the output

(1) Beta Distribution

- **Estimation method**

  The pdf of the Beta Distribution is given as following:

$$f(x|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

$$0 \le x \le 1, \alpha > 0, \beta > 0$$

  The Likelihood Function for this is as follows:

$$L(x_i|\alpha,\beta) = \prod_{i=1}^{n} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x_i^{\alpha-1}(1-x_i)^{\beta-1}$$

  The value of α and β that maximize this function are known as the Maximum Likelihood Estimators. To simplify the above equation by taking the logarithm,

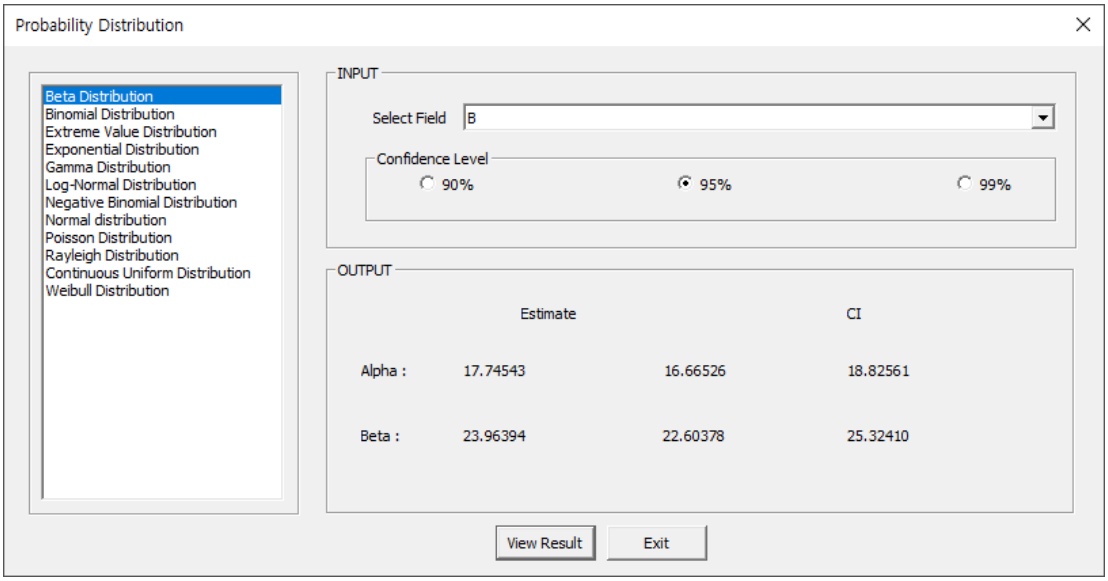
$$ln\, L(x_i|\alpha,\beta) = n\, ln\, \Gamma(\alpha+\beta) - n\, ln\, \Gamma(\alpha) - n\, ln\, \Gamma(\beta) + (\alpha - 1)\sum_{i=1}^{n} ln\, x_i + (\beta$$

$$- 1)\sum_{i=1}^{n} ln(1-x_i)$$

  There are several methods to maximize the above equation. ECMiner™ employs Nelder and Mead's Simplex method to maximize the Likelihood Function.

  Caution: All data used in Beta Distribution must be between 0 and 1.

- **Example**



(2) Binomial Distribution

- **Estimation method**

The pmf of the Binomial Distribution is given as following:

$$f(x|n,p) = \binom{n}{x} p^x (1-p)^{n-x}$$

The Likelihood Function for this is as follows:

$$L(x_i|n,p) = \prod_{i=1}^{n} \binom{n}{x_i} p^{x_i} (1-p)^{n-x_i}$$

The value of n and p that maximize this function are known as the Maximum Likelihood Estimators. To simplify the above equation by taking the logarithm,

$$ln\, L(x_i|n,p) = ln \binom{n}{x_i} + x_i\, ln\, p + (n - x_i)\, ln(1-p)$$

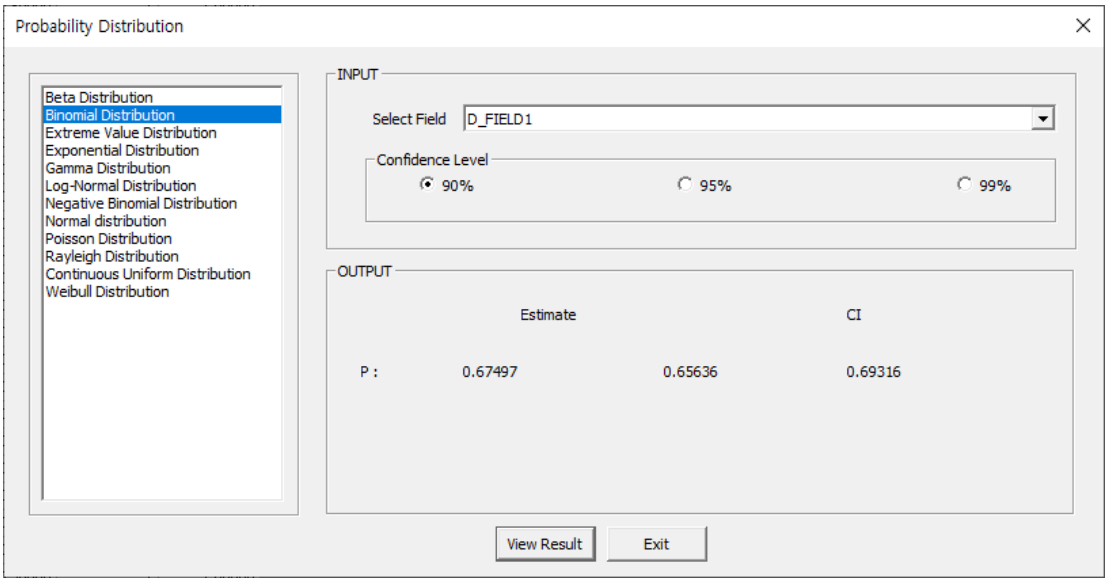
Take the partial derivative of the above equation and set it equal to zero to maximize the likelihood function.

$$\frac{\partial}{\partial p} ln\, L(x_i|n,p) = \frac{x_i}{p} - \frac{n - x_i}{1-p} = 0$$

The maximum likelihood estimate is as follows.

$$\hat{p} = \frac{x_i}{n}$$

- **Example**



---

(4) Exponential Distribution

- **Estimation method**

pdf of the Exponential Distribution is given as follows:

$$f(x|\lambda) = \lambda e^{-\lambda x}$$

The corresponding Likelihood Function is:

$$L(\lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i}$$

The value of λ that maximize this function are known as the Maximum Likelihood Estimators. To simplify the above equation by taking the logarithm,

$$ln\, L(\lambda) = \sum_{i=1}^{n} ln(\lambda) - \lambda \sum_{i=1}^{n} x_i$$

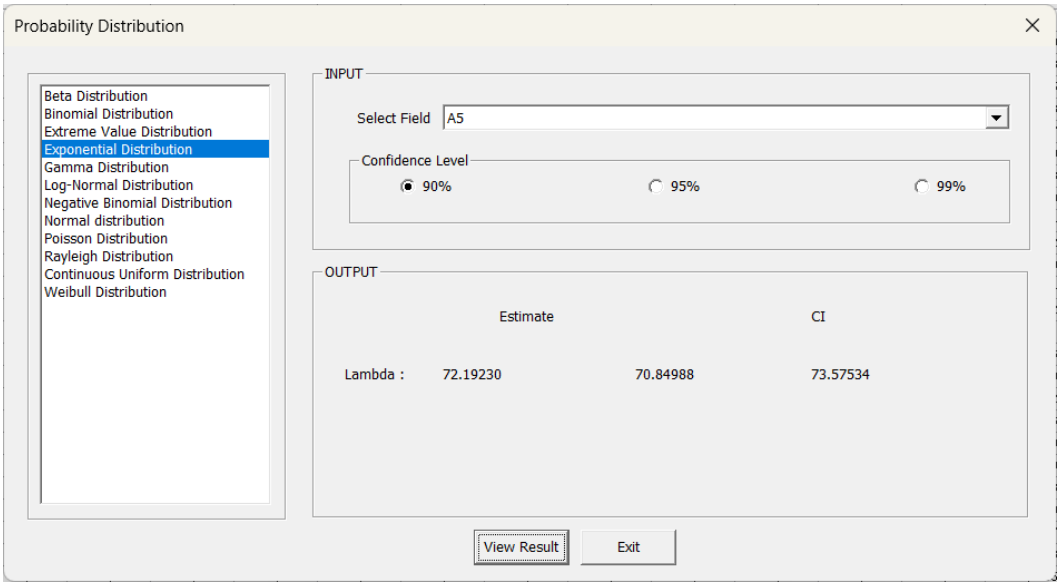
Take the partial derivative of the above equation and set it equal to zero to maximize the likelihood function.

$$\frac{\partial}{\partial \lambda} ln\, L(\lambda) = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i = 0$$

Looking at the above, the maximum likelihood estimate is as follows.

$$\hat{\lambda} = \frac{n}{\sum_{i=1}^{n} x_i}$$

- **Example**



---

(5) Gamma Distribution

- **Estimation method**

The pdf of the Gamma Distribution is as follows:

$$f(x|\alpha, \beta) = \frac{1}{\beta^{\alpha}\Gamma(\alpha)} x^{\alpha-1} e^{-\frac{x}{\beta}}$$

$$x > 0, \quad \alpha > 0, \quad and \quad \beta > 0$$

The Likelihood Function for this is as follows:

$$L(x_i|\alpha, \beta) = \prod_{i=1}^{n} \frac{1}{\beta^{\alpha}\Gamma(\alpha)} x_i^{\alpha-1} e^{-\frac{x_i}{\beta}}$$

The value of α and β that maximize this function are known as the Maximum Likelihood Estimators. To simplify the above equation by taking the logarithm,

$$ln\,L(x_i|\alpha,\beta) = -n\alpha\,ln\,\beta - n\,ln\,\Gamma(\alpha) + (\alpha-1)\sum_{i=1}^{n} ln\,x_i - \frac{1}{\beta}\sum_{i=1}^{n} x_i$$

Take the partial derivative of the above equation and set it equal to zero to maximize the likelihood function.

$$\frac{\partial}{\partial\alpha} ln\,L(x_i|\alpha,\beta) = -n\,ln\,\beta + \sum_{i=1}^{n} ln\,x_i - n\frac{\partial\,ln\,\Gamma(\alpha)}{\partial\alpha} = 0$$

By using Newton Raphson's method:

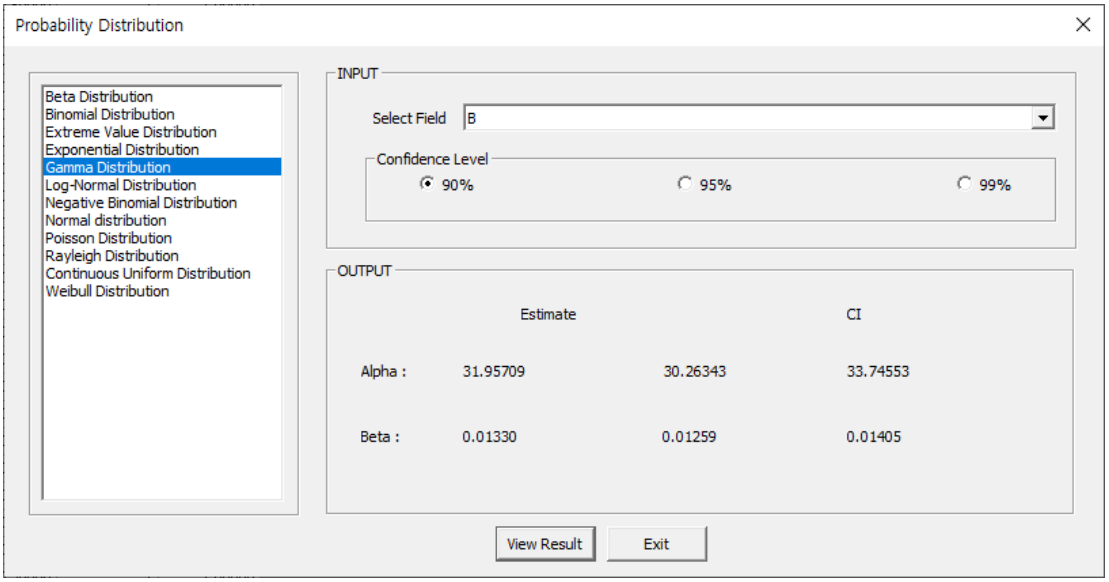
$$\hat{\alpha} = \alpha - \frac{f(\alpha)}{f'(\alpha)}$$

$$where\ f'(\alpha) = -n\frac{\partial^2\,ln\,\Gamma(\alpha)}{\partial\alpha} + \frac{n}{\alpha}$$

$$\frac{\partial}{\partial\beta} ln\,L(x_i|\alpha,\beta) = -\frac{n\alpha}{\beta} + \frac{1}{\beta^2}\sum_{i=1}^{n} x_i = 0$$

$$\hat{\beta} = \frac{\overline{x}}{\alpha}$$

- **Example**



---

(8) Normal Distribution

- **Estimation method**

The pdf of Normal Distribution is given as following:

$$\boldsymbol{f(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}}$$

The corresponding Likelihood Function is:

$$\boldsymbol{L(x_i|\mu,\sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{\sigma^2}}}$$

The value of μ and σ^2that maximize this function are known as the Maximum Likelihood Estimators. To simplify the above equation by taking the logarithm,

$$\boldsymbol{ln\, L(x_i|\mu,\sigma^2) = \frac{n}{2} ln\, 2\pi\ - \frac{n}{2} ln\, \sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}$$

Take the partial derivative of the above equation and set it equal to zero to maximize the likelihood function.

$$\boldsymbol{\frac{\partial}{\partial\mu} ln\, L(x_i|\mu,\sigma^2) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 = 0}$$

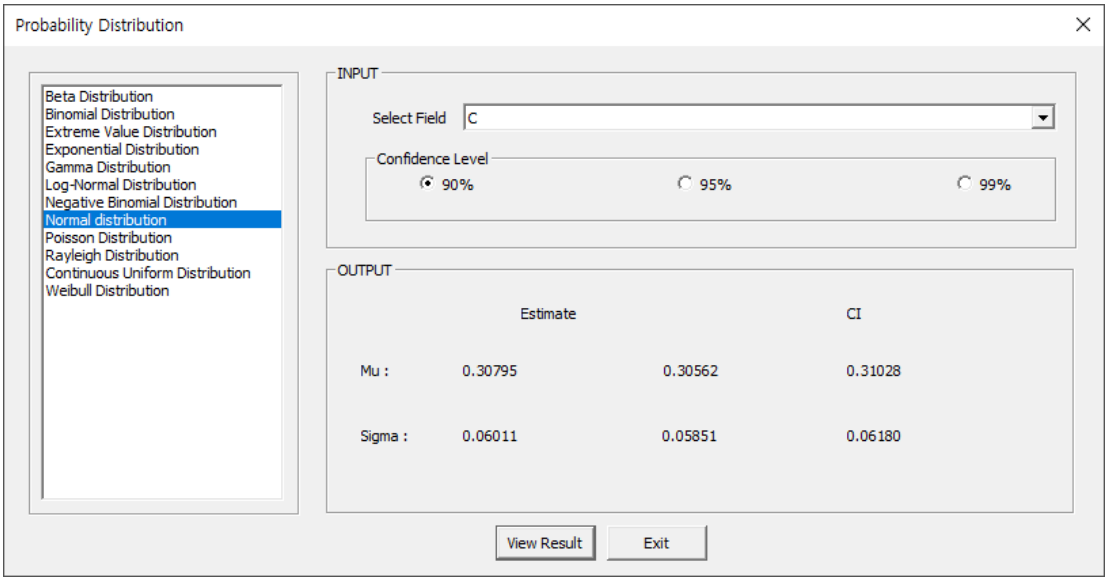
$$\boldsymbol{\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i}$$

$$\boldsymbol{\frac{\partial}{\partial\sigma^2} ln\, L(x_i|\mu,\sigma^2) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i-\mu)^2 = 0}$$

$$\boldsymbol{\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(x_i-\widehat{\mu})^2}$$

Here X_α is the α quantile of Normal Distribution.

- **Example**



(9) Poisson Distribution

- **Estimation method**

The pdf of Poisson Distribution is given as following:

$$f(x|\lambda) = \frac{\lambda^x}{x!}e^{-\lambda}$$

The corresponding Likelihood Function is:

$$L(x_i|\lambda) = \prod_{i=1}^{n} \frac{\lambda^x}{x!}e^{-\lambda}$$

The value of λ that maximize this function are known as the Maximum Likelihood Estimators. To simplify the above equation by taking the logarithm,

$$ln\, L(x_i|\lambda) = \sum_{n=1}^{n} ln\left(\frac{\lambda^{x_i}}{x_i!}e^{-\lambda}\right)$$

$$= -n\lambda + \sum_{i=1}^{n} x_i\, ln\, \lambda - \sum_{i=1}^{n} x_i\,!$$

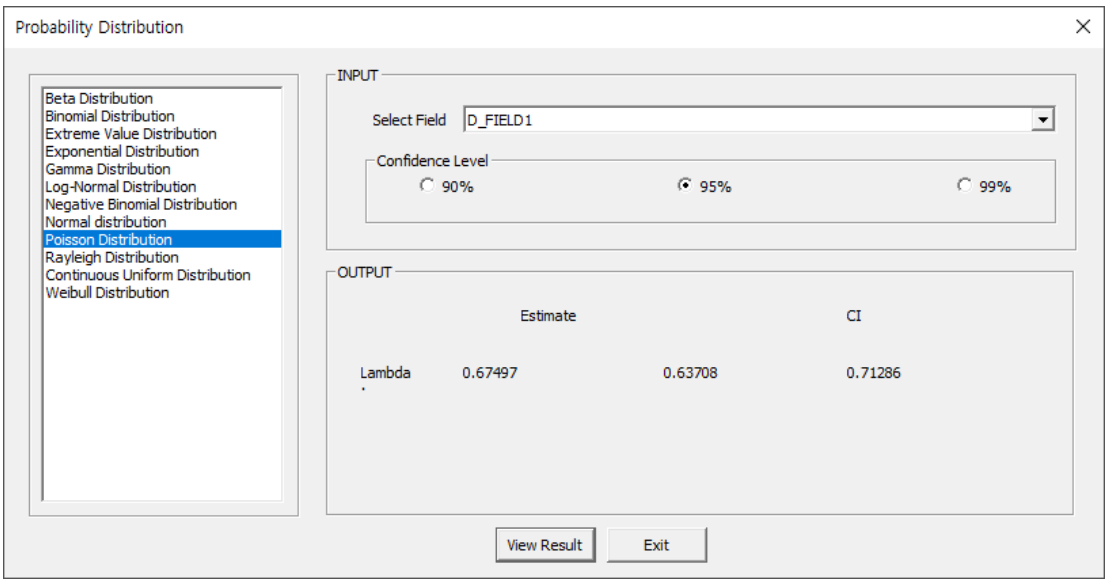
Take the partial derivative of the above equation and set it equal to zero to maximize the likelihood function.

$$\frac{\partial}{\partial \lambda} ln\, L(x_i|\lambda) = -n + \frac{1}{\lambda}\sum_{i=1}^{n} x_i\, ! = 0$$

Solving this yields the following estimates:

$$\hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- **Example**



---

(10) Continuous Uniform Distribution

- **Estimation method**

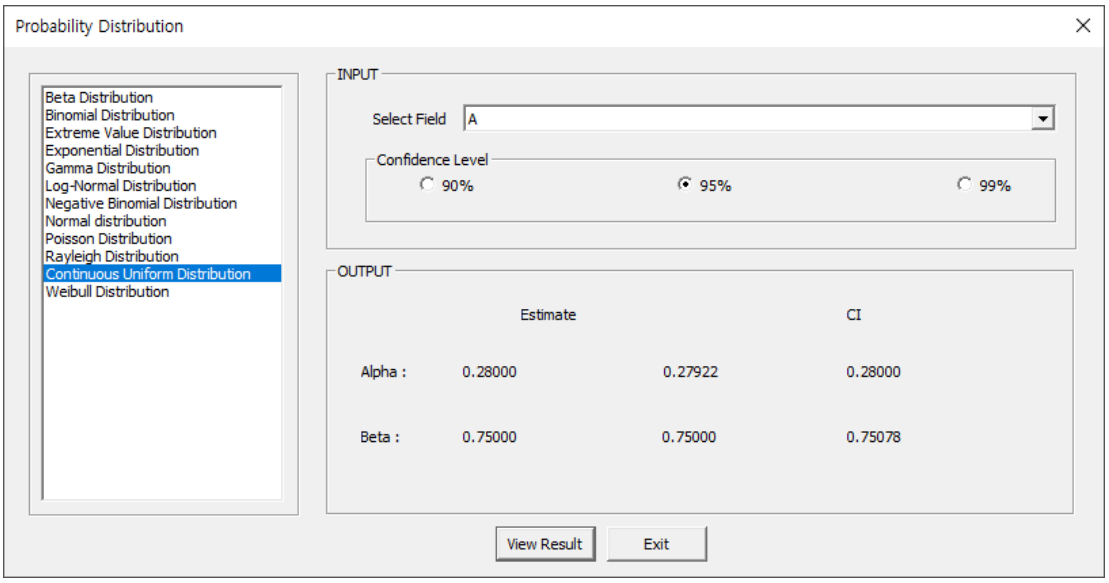
The pdf of Continuous Uniform Distribution is given as follows:

$$f(x|a,b) = \frac{1}{b-a}$$

And the MLE for a and b are:

$$\hat{a} = min(x_1, x_2, .., x_n)$$
$$\hat{b} = max(x_1, x_2, .., x_n)$$

- **Example**



---

(11) Weibull distribution

- **Estimation method**

The pdf of the Weibull distribution is given as follows:

$$f(x|\alpha,\beta) = \left(\frac{\beta}{\alpha}\right)\left(\frac{x}{\alpha}\right)^{\beta-1} e^{-\left(\frac{x}{\alpha}\right)^{\beta}}$$

$$x \geq 0, \alpha > 0, \beta > 0$$

he corresponding Likelihood Function is:

$$L(x_i|\alpha,\beta) = \prod_{i=1}^{n}\left(\frac{\beta}{\alpha}\right)\left(\frac{x_i}{\alpha}\right)^{\beta-1} e^{-\left(\frac{x_i}{\alpha}\right)^{\beta}}$$

The value of λ that maximize this function are known as the Maximum Likelihood Estimators. To simplify the above equation by taking the logarithm,

$$ln\, L(x|\alpha,\beta) = n\, ln\, \beta - n\, ln\, \alpha + (\beta-1)\sum_{i=1}^{n} ln\, x_i - \sum_{i=1}^{n}\left(\frac{x_i}{\alpha}\right)^{\beta}$$

Take the partial derivative of the above equation and set it equal to zero to maximize the likelihood function.

$$\frac{\partial}{\partial \alpha} ln\, L(x_i|\alpha,\beta) = -\frac{n}{\alpha} + \frac{\beta}{\alpha^2}\sum_{i=1}^{n} x_i^{\beta} e^{-\left(\frac{x_i}{\alpha}\right)^{\beta}} = 0$$

$$\frac{\partial}{\partial \beta} ln\, L(x_i|\alpha,\beta) = \frac{n}{\beta} + \sum_{i=1}^{n} ln\, x_i - \sum_{i=1}^{n} \left(\frac{x_i}{\alpha}\right)^{\beta} ln\left(\frac{x_i}{\alpha}\right) = 0$$

Solving this yields the following estimates:

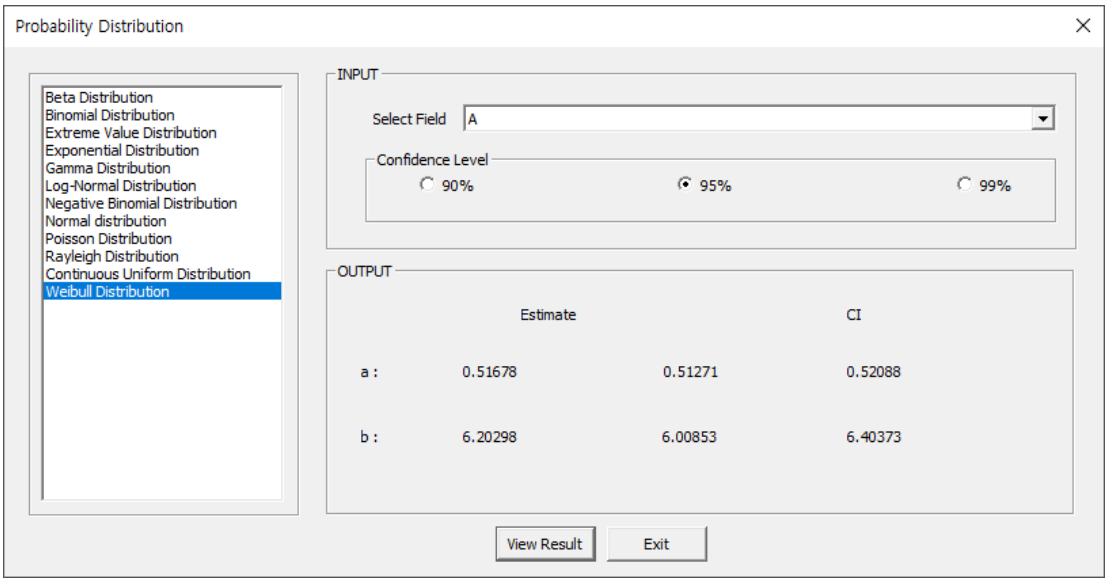
$$\hat{\alpha} = \left(\frac{1}{n}\sum_{i=1}^{n} x_i^{\beta}\right)^{\frac{1}{\beta}}$$

By using Newton – Raphson method β solved as

$$\beta_{\beta+1} = \beta_{\beta} - \frac{f(\beta_{\beta})}{f'(\beta_{\beta})}$$

$$where\ f'(\beta_{\beta}) = \frac{1}{n}\left(\sum_{i=1}^{n} ln\, x_i\right)\left(\sum_{i=1}^{n} x_i^{\beta} - \beta\sum_{i=1}^{n} x_i^{\beta}\, ln\, x_i\right) - \beta\sum_{i=1}^{n} x_i^{\beta} (ln\, x_i)^2$$

- **Example**

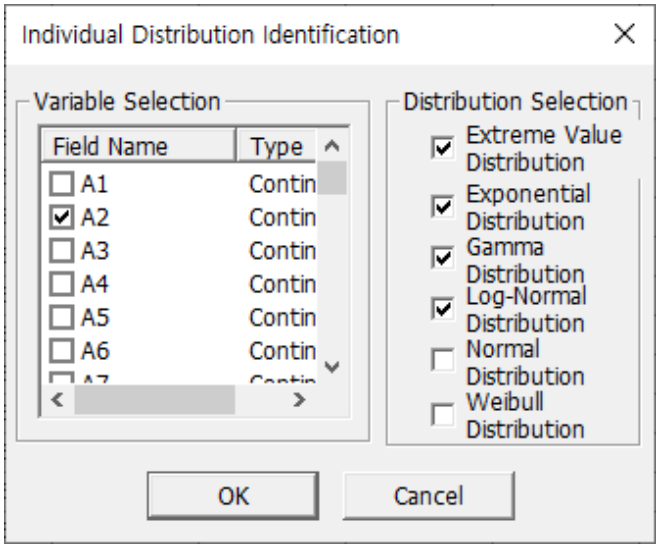


4.3.8.2 Individual Distribution Identification

To check the distribution of data, one can apply various distributions and utilize the obtained Anderson-Darling statistic along with its P-value to determine the most similar distribution to the data.

How to run

[Analyze] – [Probability Distribution] – [Individual Distribution Identification]

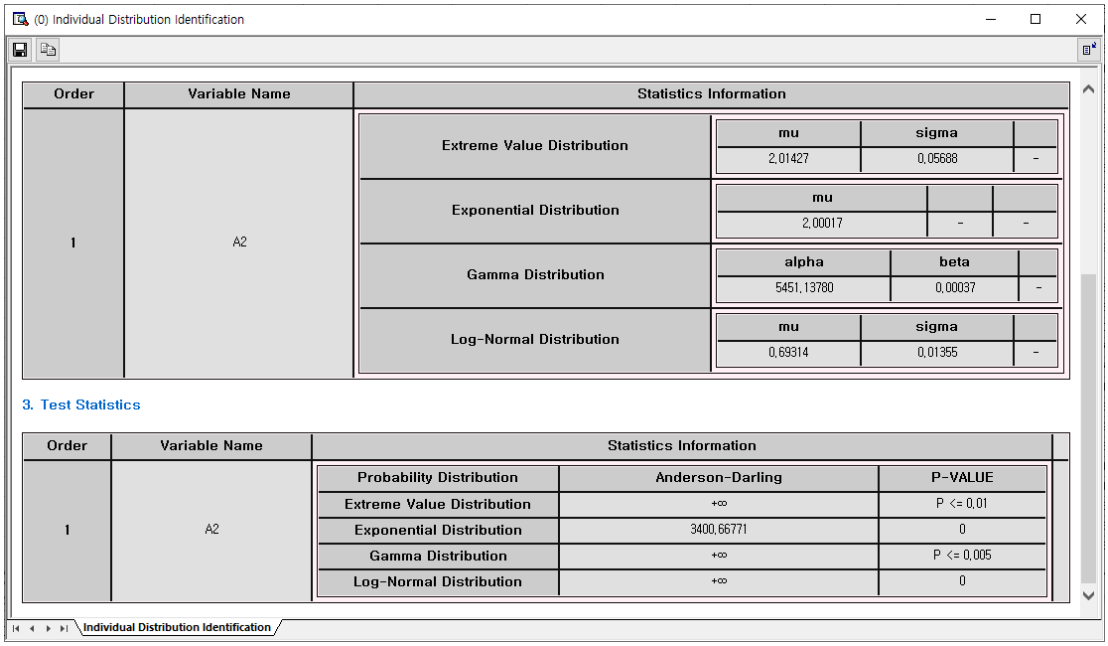


From the variable list, select the variable for which distribution identification is desired (multiple selections are allowed), and then choose which distributions to apply from the distribution selection (multiple selections are allowed).

Result

- **Variable information**: Provides basic statistics of the selected variable(s).
- **Distribution Parameter Estimation**: Estimates the parameters of the selected distribution based on the relevant variable.
- **Test statistic**: Provides the Anderson-Darling statistic and P-value, indicating how well the variable fits the chosen distribution.

  Note: If the P-value exceeds 0.05, it can be concluded that variable adheres to the specified distribution.



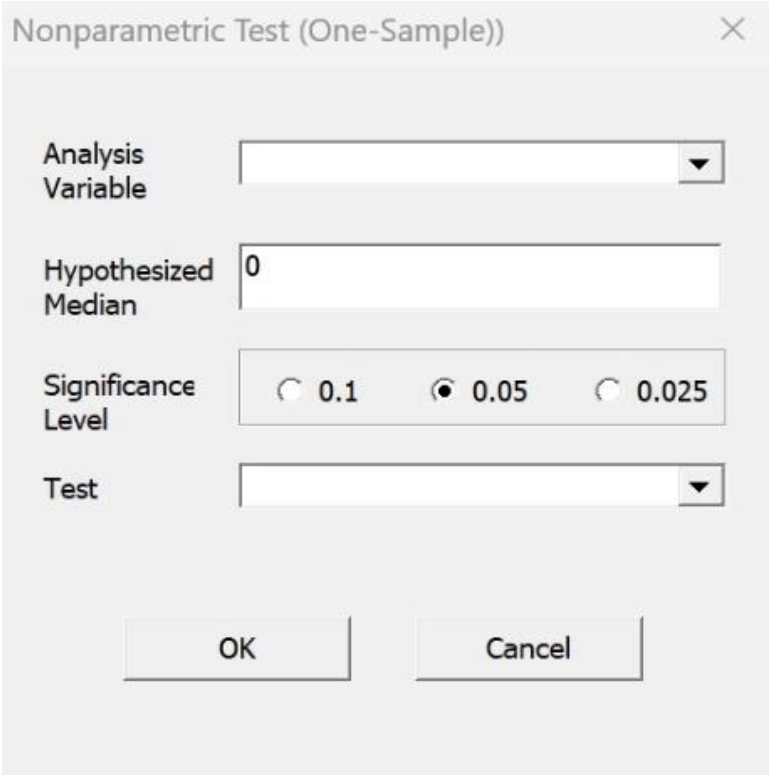
4.3.9 Nonparametric Test

4.3.9.1 One-Sample

A non-parametric one-sample test is used to evaluate whether the median of a sample differs from a hypothesized value. It does not assume a specific distribution and is useful when the data is not normally distributed. Our software supports the Wilcoxon signed-rank test, which takes into account the rank of values.

How to run

[Analyze] – [Nonparametric Test]



Select Analysis **Variable** and enter the **hypothesized median**. Set the **Significance level**. Choose **one** of two testing methods.

Result

One-Sample Nonparametric Test

One-Sample Statistics

Variable Name	Data Count	Median	Average	Standard Deviation	Mean Standard Error
A1	7596	109,99100	109,35402	1,74303	0,02000

One-Sample Test

Null Hypothesis	Hypothesis Testing Method	Significant Probability (Two-Sided)	Conclusion
The median value of A1 is 0,000000.	Sign Test	0	Reject the null hypothesis.

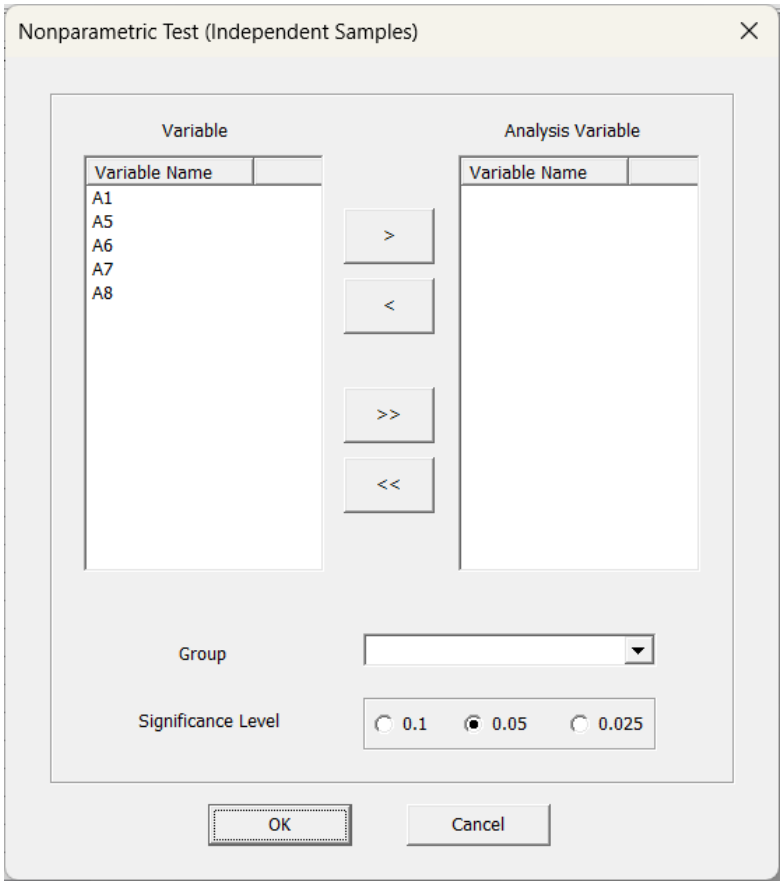
* Used normal distribution approximation when calculating test statistics.

4.3.9.2 Independent Samples

Non-parametric independent samples tests are used to compare two or more groups without assuming a specific distribution. Our software provides the Mann-Whitney U statistic method.

How to run

[Analyze] – [Nonparametric Test] - [Independent Samples]



Select **Analysis Variable**, and **Group variable**. Set the **Significance level**.

Result

Nonparametric Test (Independent Samples)

Group statistics

Nonparametric Independent Samples Test

	churn_status	Number of Data Points	Average	Standard Deviation	Standard Error of the Mean
A1	0	5	0	0	0
	1	19	17,73684	-30,01754	-6,88650

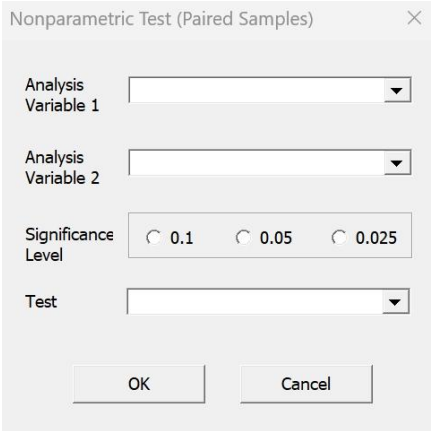
Null Hypothesis	Hypothesis Testing Method	Significant Probability (Two-Sided)	Conclusion
The distribution of A1 is identical regardless of the group of churn_status.	Mann-Whitney U Test	0,00001	Reject the null hypothesis.

4.3.9.3 Paired Samples

For non-parametric paired samples, there are two tests, the sign test and Wilcoxon signed-rank test.

How to run

[Analyze] – [Nonparametric Test] - [Paired Samples]



Select two variables, **variable1** and **variable2**. Set the **Significance level**.

Result

Nonparametric Test (Paired Samples)

Paired Sample Statistics

	Average	Number of Data Points	Median	Standard Deviation	Standard Error of the Mean
A1	29,37500	24	24	13,03778	2,66133
A5	145,48341	24	146,98750	43,72323	8,92497

Paired Sample Statistics

Null Hypothesis	Hypothesis Testing Method	Significant Probability(Two-Sided)	Conclusion
The median difference between A1 and A5 is 0.	Wilcoxon Signed Rank Test	0,00002	Reject the null hypothesis.

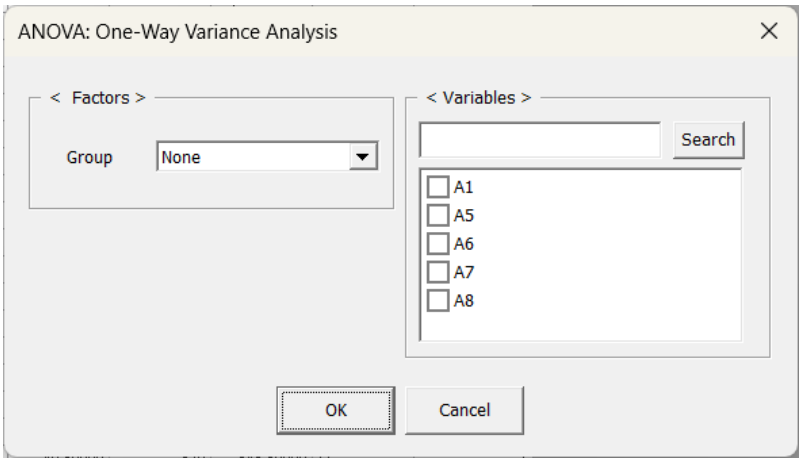
* Used normal distribution approximation when calculating test statistics.

4.3.9.4 ANOVA- One-way

The Kruskal-Wallis test is a non-parametric method used to compare the medians of three or more independent groups. It's an alternative to one-way ANOVA for non-normally distributed data.

How to run

[Analyze] – [Nonparametric Test] - [ANOVA-One-way]



Specify the **group** as **Factors** and select **Variables.**

Result

Nonparametric Test (One-way ANOVA)

Observed Variable Name: A1

Null Hypothesis	Testing Method	Freedom Degree	Hypothesis Test Statistics	Significant Probability	Conclusion
The distribution of A1 is identical regardless of the group of Group.	Kruskal-Wallis Test	1	1,83048	0,17607	Do not reject the null hypothesis.

Group statistics

Level	N	Mean	Median	Ave Rank	StdDev
M	21	30,52381	24	13,23810	12,85679
MH	3	21,33333	18	7,33333	8,49837

4.3.10 Accuracy Measurement

The model's accuracy and validity can be measured by comparing the true(original) values with the predicted values.

How to run

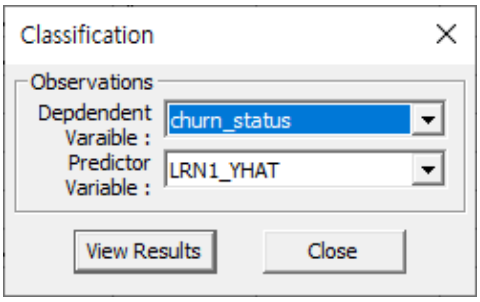
[Classification] or [Regression] from the submenu of [Analyze] – [Accuracy Measurement].

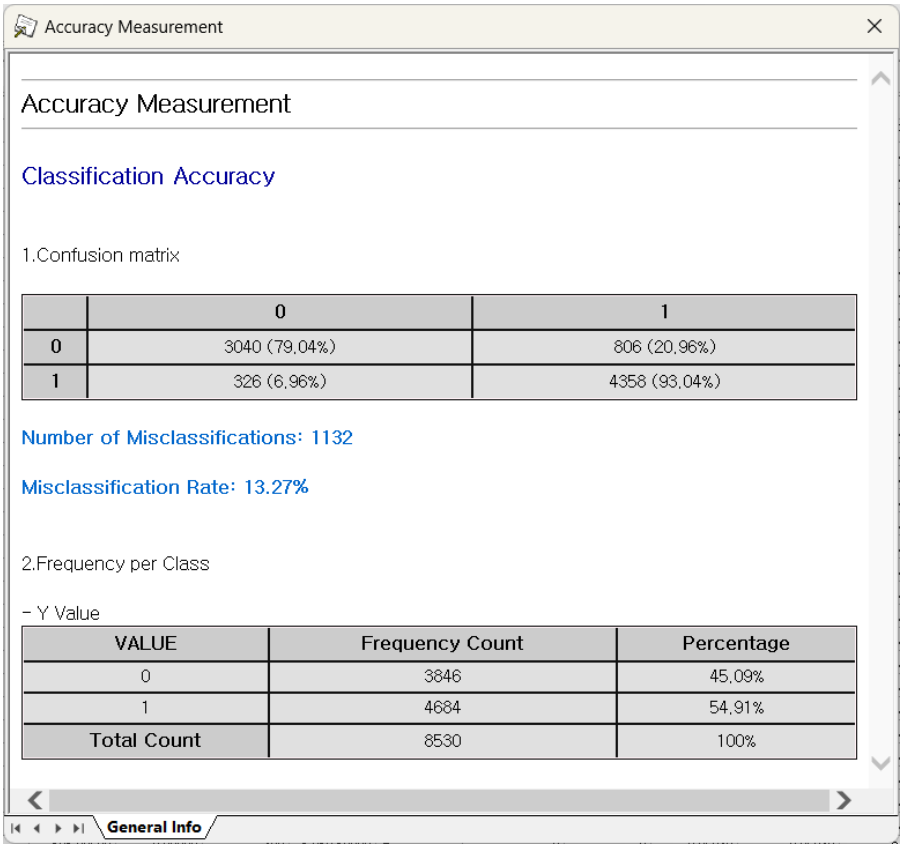
4.3.10.1 Classification

- **Estimation method**

  The **classification** analysis predicts each class label. A **confusion matrix** is a table used to evaluate the performance of a classification model. **Confusion matrix** is a 2x2 table with the frequency of true tartet class and predicted class. It shows true positive rate, true negative rate, false positive rate, and false negative rate.

- **Example**





4.3.10.2 Regression

- **Estimation method**

  R-square, Mean Absolute Percentage Error (MAPE), Mean Absolute Deviation (MAD), and Mean Squared Deviation (MSD)

  (1) R-square

R-square **is a statistical measure to evaluate how well a regression model fits the data. It represents the proportion of the variance in the dependent variable. A value closer to 1 indicates better fitting model.**

(2) MAPE (Mean Absolute Percentage Error)

$$MAPE = \frac{\sum_{i=1}^{n} | (y_i - \hat{y}_i) / y_i |}{n} \times 100$$

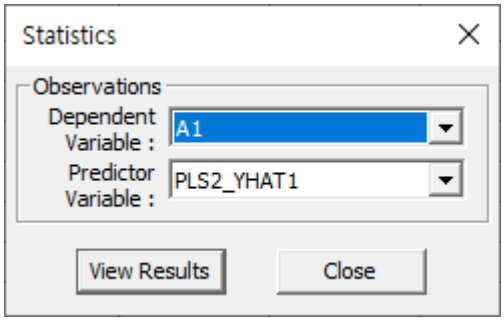
(3) MAE (Mean Absolute Error)

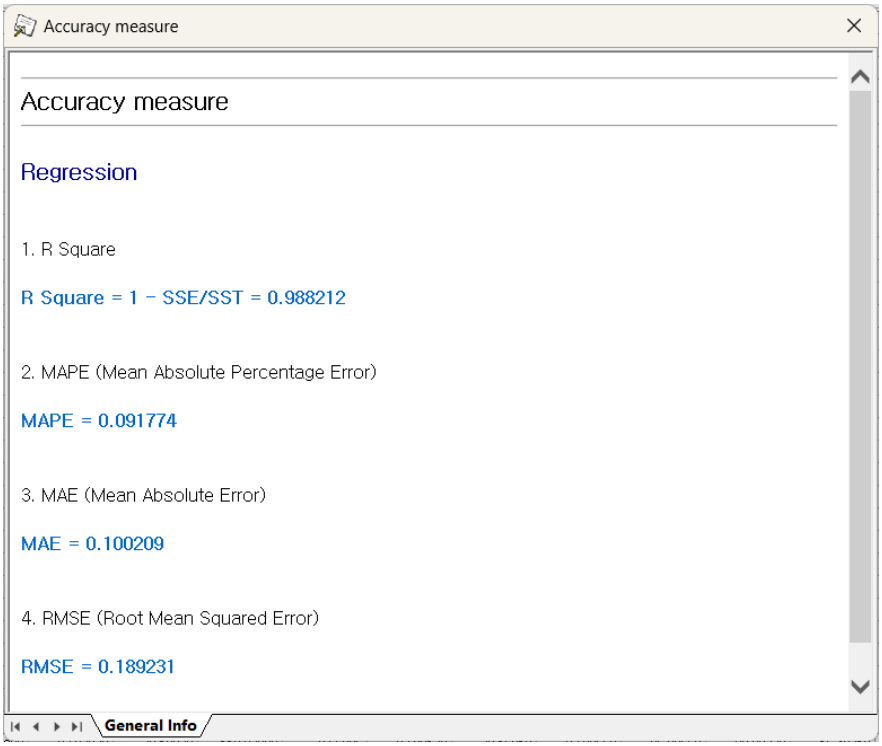
$$MAE = \frac{\sum_{i=1}^{n} | y_i - \hat{y}_i |}{n}$$

(4) RMSE (Root Mean Squared Error)

$$RMSE = \sqrt{MSE} = \sqrt{SSE/(n-p-1)}$$

- **Example**





4.3.11 Gage R&R

Gage R&R (Gage Repeatability and Reproducibility) is a statistical method used to evaluate the amount of variation in a measurement system arising from the measurement device (repeatability) and the people taking the measurements (reproducibility).

4.3.11.1 Gage Run Chart

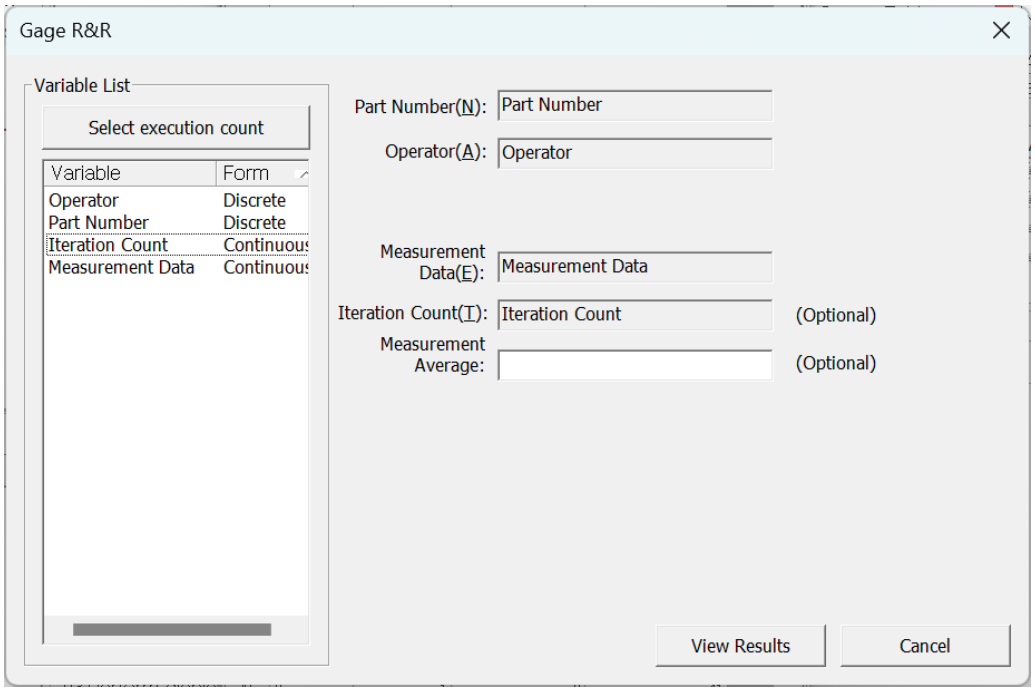
Gage Run Chart is a visual tool used to assess the consistency of a measurement system by plotting measurement values based on **Part Number**, **Operator**, **Measurement Data**, and **Iteration Count**. The chart displays the measurements taken by different operators over several iterations and compares them against a **Measurement Average** or **calculated mean** to evaluate the stability and variability of the measurement system over time.

An example of the data used is shown below.

	1	2	3	4
	Operator	Part Number	Iteration Count	Measurement Data
1	A	1	1	37
2	A	2	1	42
3	A	3	1	30
4	A	4	1	42
5	A	5	1	28
6	A	6	1	42
7	A	7	1	25
8	A	8	1	40
9	A	9	1	25
10	A	10	1	35
11	A	1	2	38
12	A	2	2	41
13	A	3	2	31
14	A	4	2	43
15	A	5	2	30
16	A	6	2	42
17	A	7	2	26
18	A	8	2	40
19	A	9	2	25
20	A	10	2	34

How to run

[Analyze] – [Gage R&R] – [Gage Run Chart]

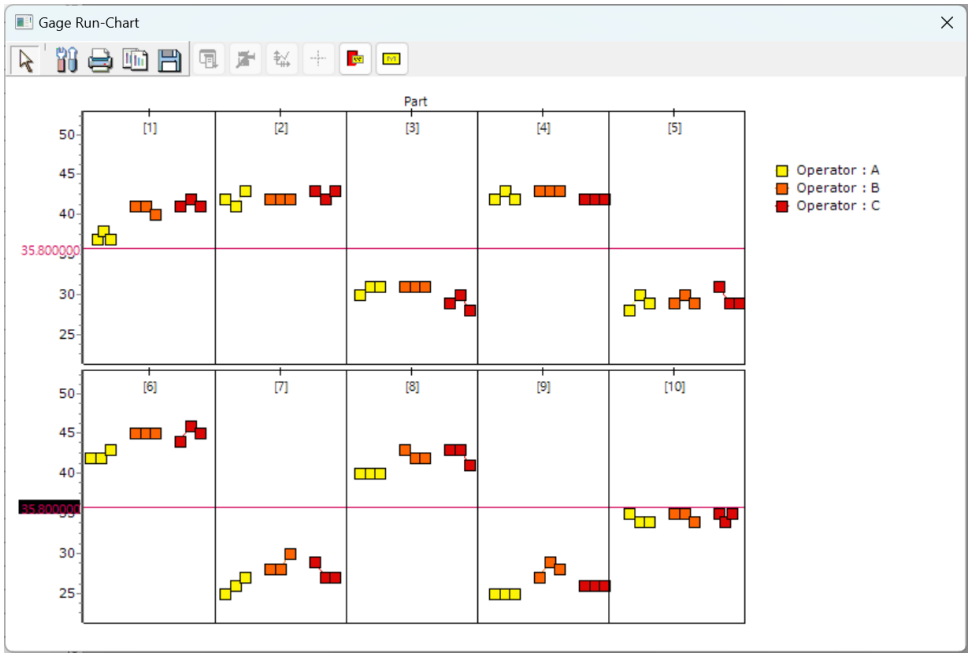


- **Part Number**: Select the parts (Must be discrete)
- **Operator**: Select the one who takes measures (Must be discrete)
- **Measurement Data**: Select the measured data (Must be continuous)
- **Iteration Count (Optional)**: Repeated measure count for the same part by the same operator. **(Must be continuous)**
- **Measurement Average (Optional)**: Enter a Measurement Average to compare the part's

measurements. By default, calculated mean of all measurements is shown.

Results

Example using 3 operators, 10 types of parts, and 3 repetitions.



4.3.11.2 Gage Linearity and Bias Study

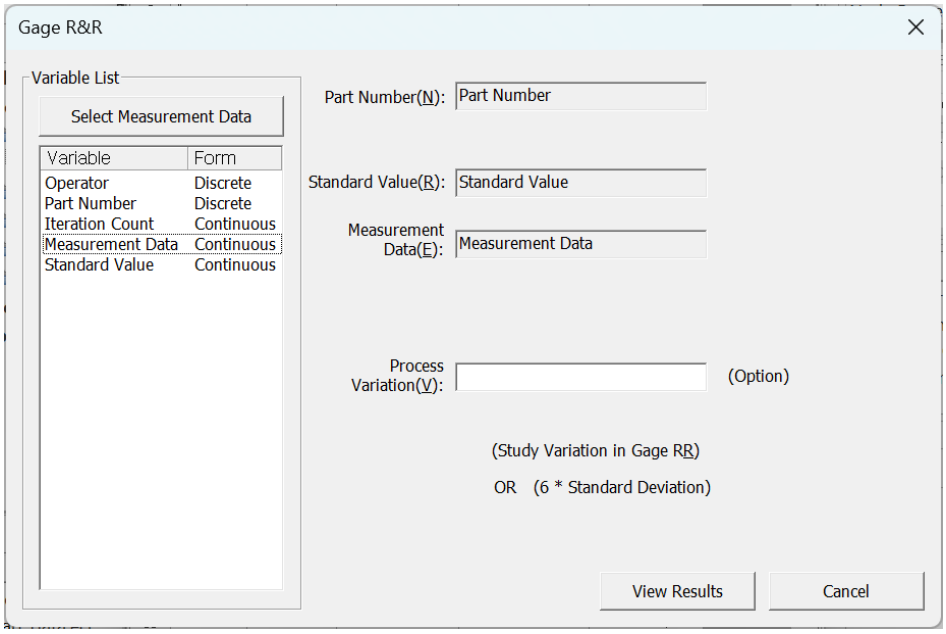
Gage Linearity and Bias Study evaluates the accuracy and consistency of the measurement system using various **Part Number**, their **Standard Value**, and **Measurement Data**. The difference between the measured values and the standard values identifies as the bias, while considering **Process Variation** to analyze whether linearity is maintained across the full measurement range.

An example of the data used is shown below.

	1	2	3	4	5
	Operator	Part Number	Iteration Count	Measurement Data	Standard Value
1	A	1	1	37	31,00000
2	A	2	1	42	32,00000
3	A	3	1	30	33,00000
4	A	4	1	42	34,00000
5	A	5	1	28	35,00000
6	A	6	1	42	36,00000
7	A	7	1	25	37,00000
8	A	8	1	40	38,00000
9	A	9	1	25	39,00000
10	A	10	1	35	40,00000
11	A	1	2	38	31,00000
12	A	2	2	41	32,00000
13	A	3	2	31	33,00000
14	A	4	2	43	34,00000
15	A	5	2	30	35,00000
16	A	6	2	42	36,00000
17	A	7	2	26	37,00000
18	A	8	2	40	38,00000
19	A	9	2	25	39,00000
20	A	10	2	34	40,00000

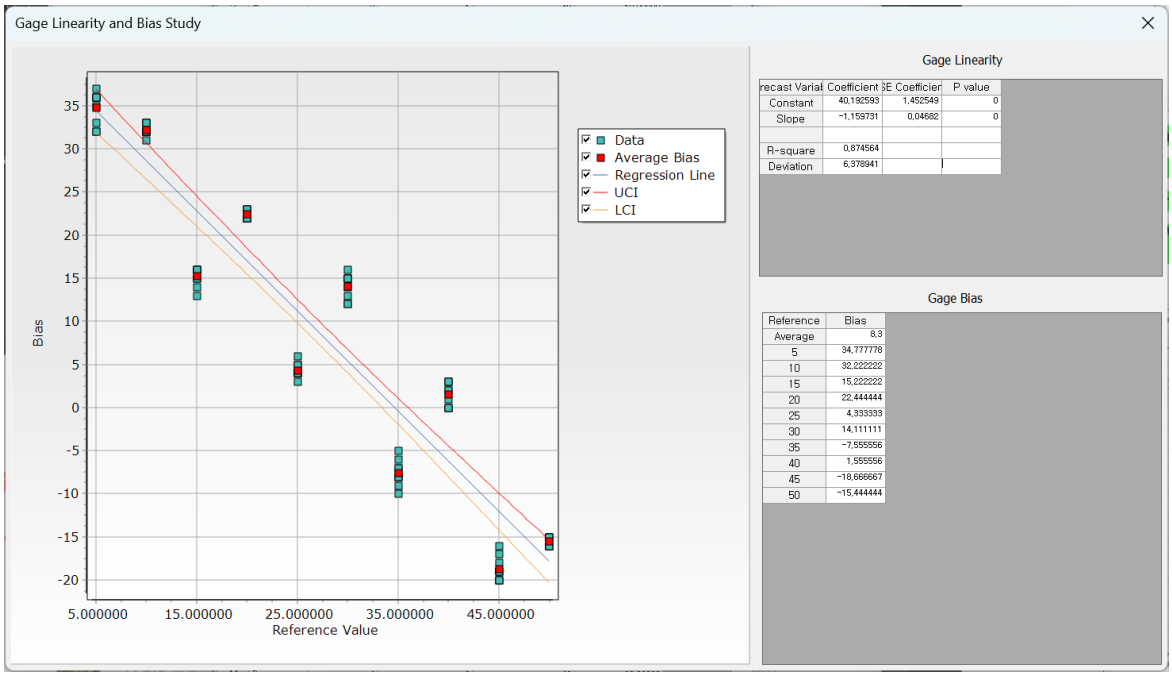
How to run

[Analyze] – [Gage R&R] – [Gage Linearity and Bias Study]



- **Part Number**: Select the parts. (Must be discrete)
- **Standard Value**: Select the standard value of the measurement (Must be continuous)
- **Measurement Data**: Select the measured data (Must be Continuous)
- **Process Variation**: Enter the process standard deviation which represents the study variation value from a Gage R&R study or 6 * standard deviation.

Results



recast Variat	Coefficient	SE Coefficien	P value
Constant	40,192593	1,452549	0
Slope	-1,159731	0,04682	0
R-square	0,874564		
Deviation	6,378941		

Reference	Bias
Average	8,3
5	34,777778
10	32,222222
15	15,222222
20	22,444444
25	4,333333
30	14,111111
35	-7,555556
40	1,555556
45	-18,666667
50	-15,444444

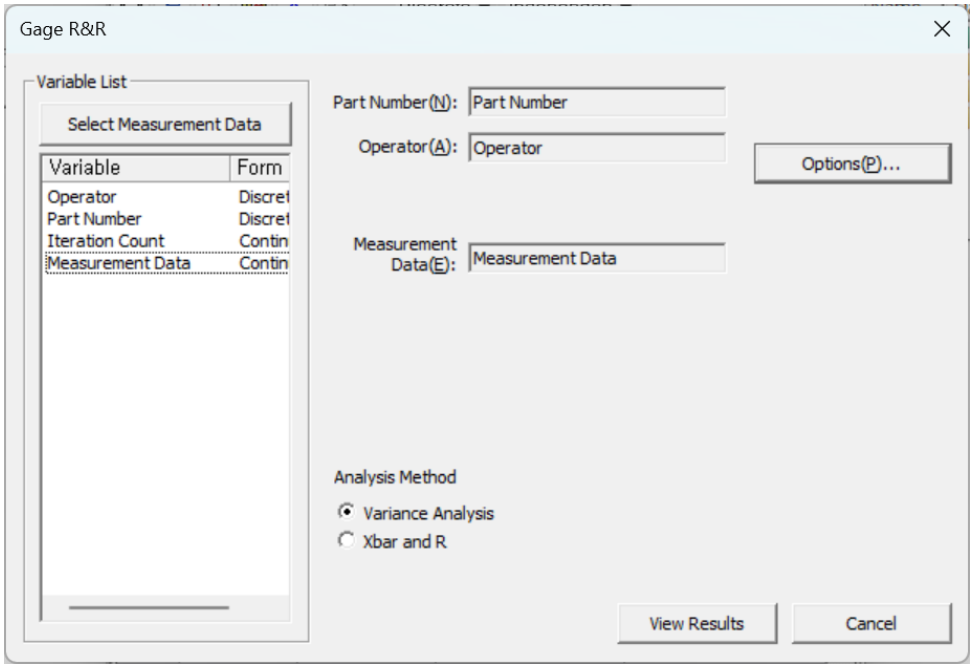
The chart on the left shows the linear relationship between bias and the data, and the table on the right provides values related to linearity and bias.

4.3.11.3 Gage R&R Study (Crossed Design)

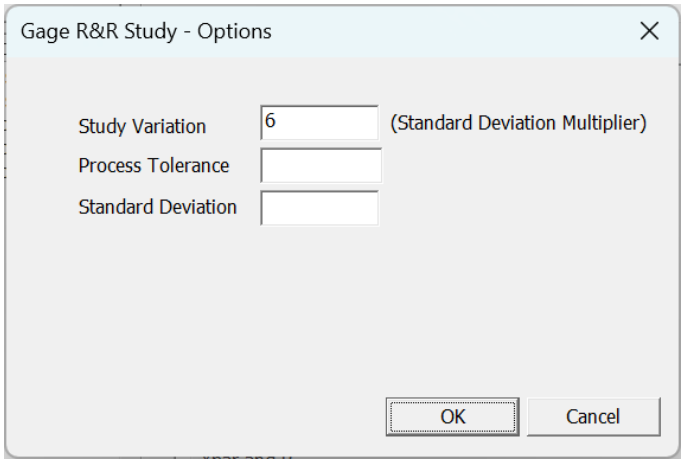
In a crossed design, every operator measures every part multiple times. It determines suitability of the measurement system. There are two methods: **Variance Analysis** and **X-bar and R** methods. If the user wants to analyze the interaction between operators and parts, a crossed design is preferable.

How to run

[Analyze] – [Gage R&R] – [Gage R&R Study (Crossed Design)]



- **Part Number**: Select the parts (Must be discrete)
- **Operator**: Select the one who takes measures (Must be discrete)
- **Measurement Data**: Select the measured data (Must be continuous)
- **Analysis Method**: Choose either **Variance analysis** or **X-bar and R** method.

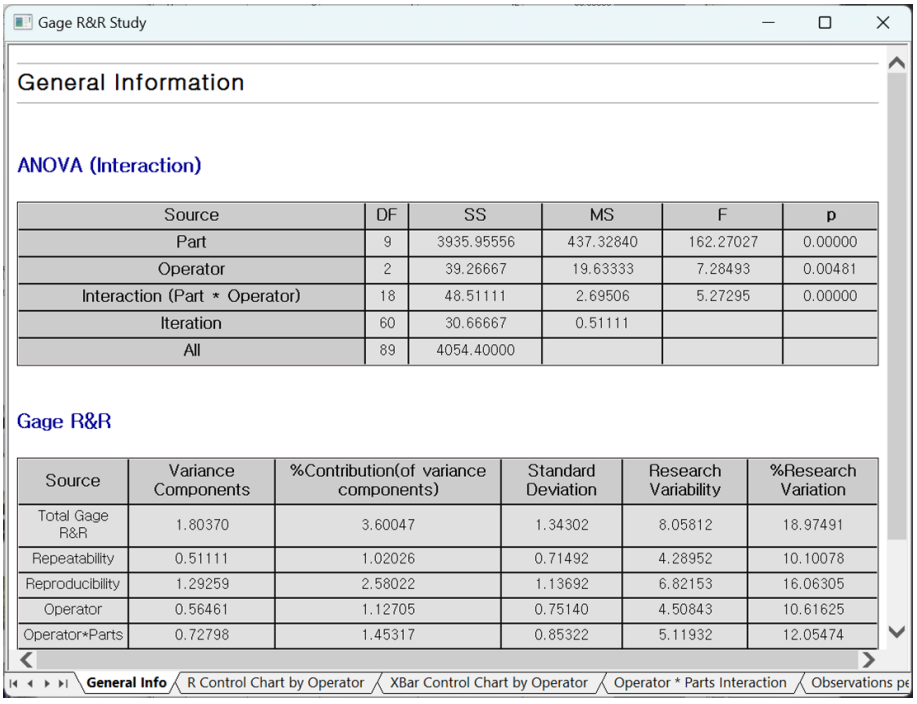


- **Study Variation**: Enter the coefficient to obtain the Study Variation.
- **Process Tolerance**: Enter an empirically known tolerance to calculate %tolerance.
- **Standard Deviation**: Enter an empirically known Historical Standard to calculate %process.

Results

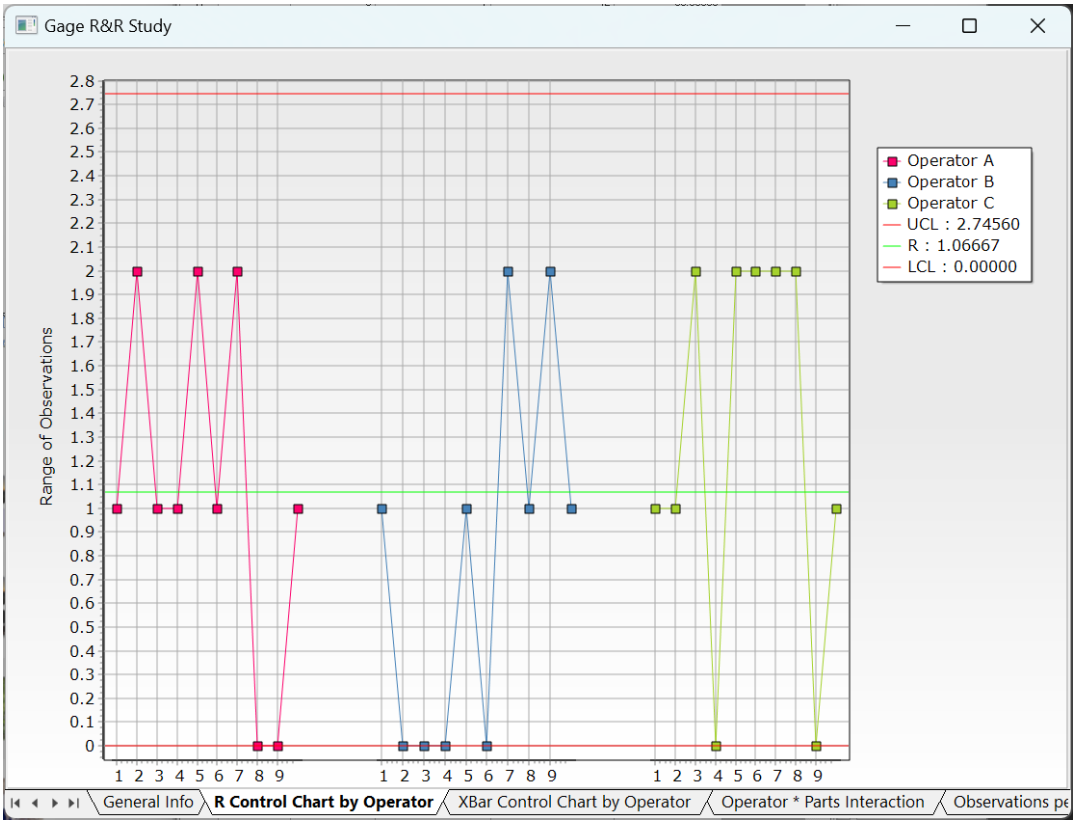
- **General Information**

  Shows the results of ANOVA (only when variance analysis is selected) and Gage R&R analysis results.



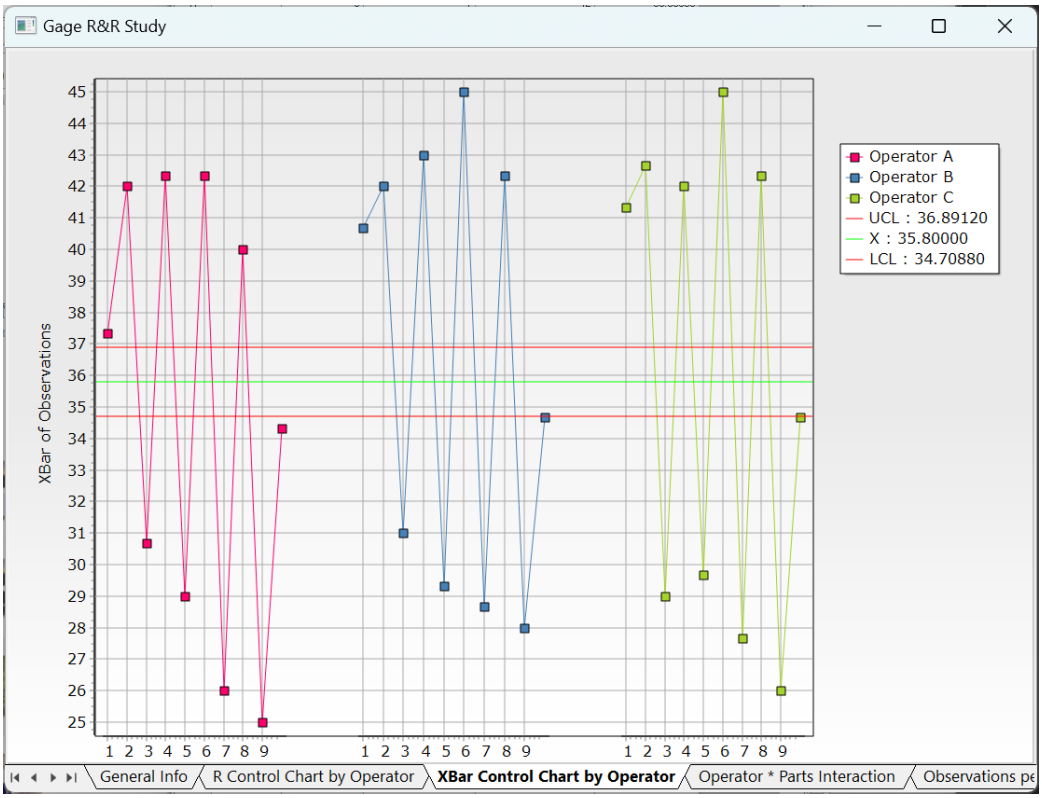
- **R Control Chart by Operator:**

  Each point displays the difference in the smallest and the biggest measurement of the same part by each operator.



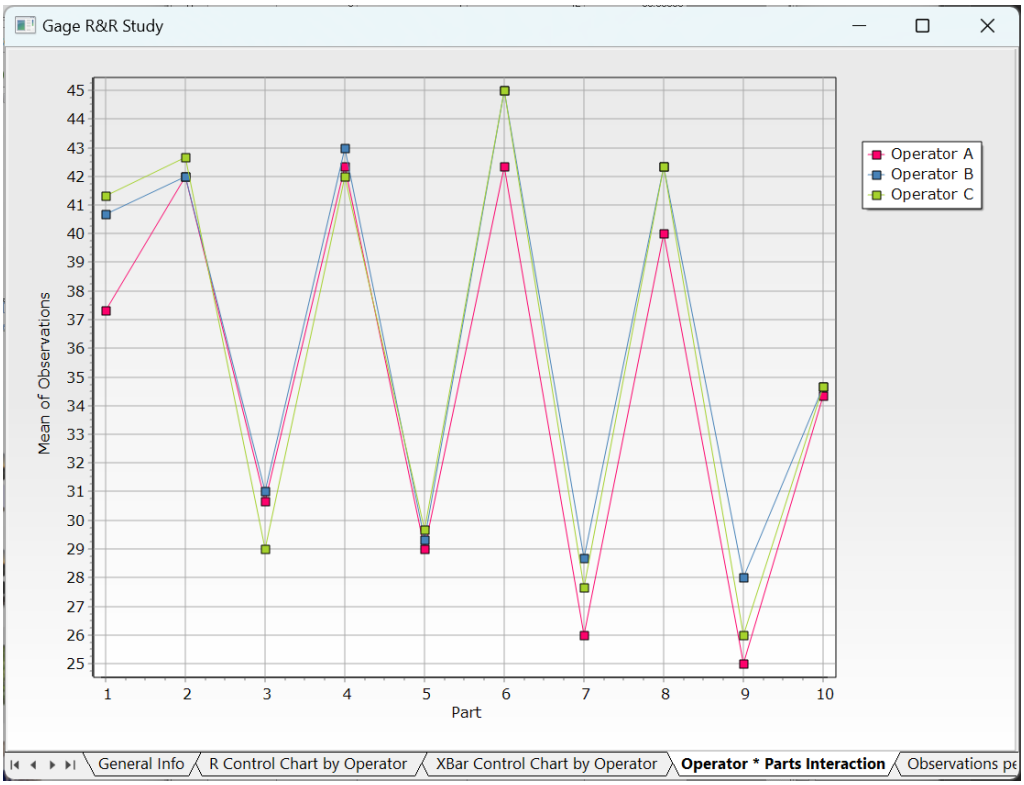
- **X-Bar Control Chart by Operator**

Each point displays the average measurement of the same part by each operator.



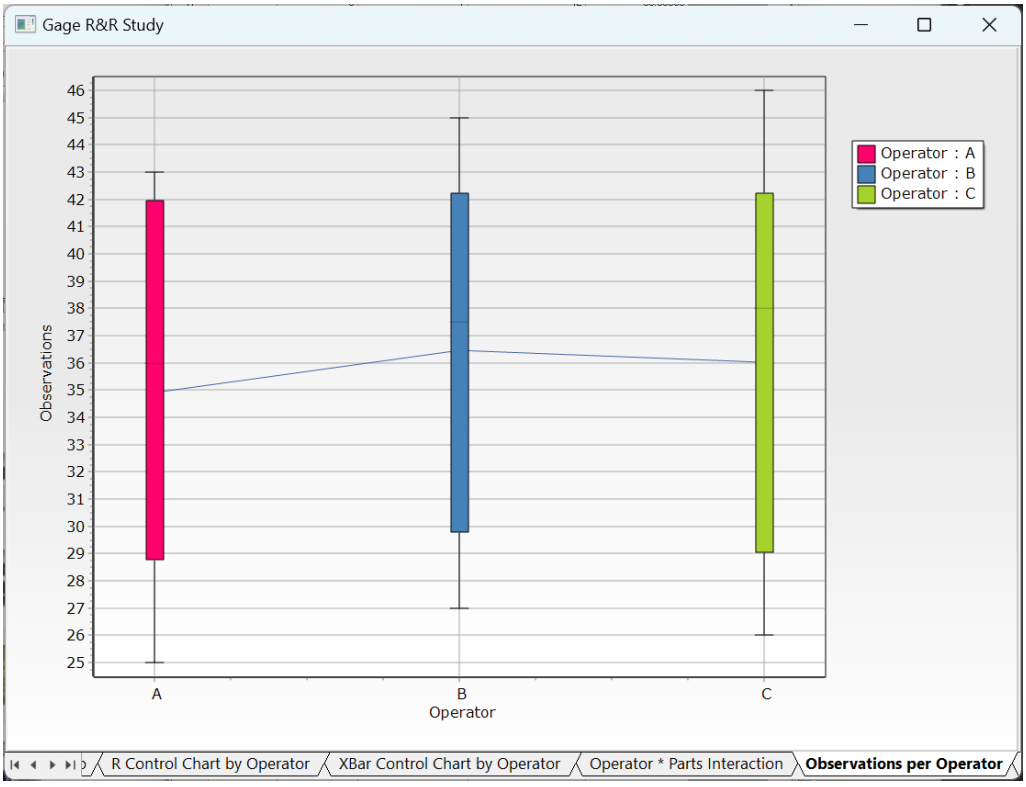
- **Operator*Parts Interaction:**

Displays How different operators interact with various parts during the measurement process.



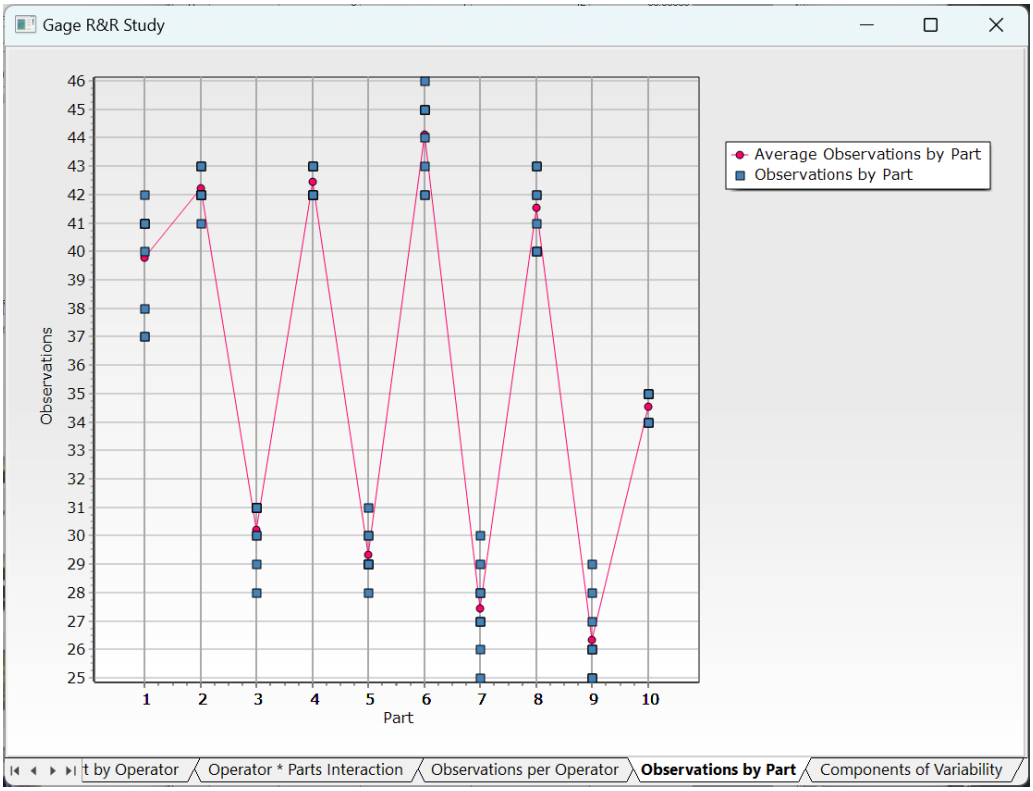
- **Observations per Operator**

  Displays box plots of all measurements for each operator, with the solid line representing the mean.



- **Observations by Part:**

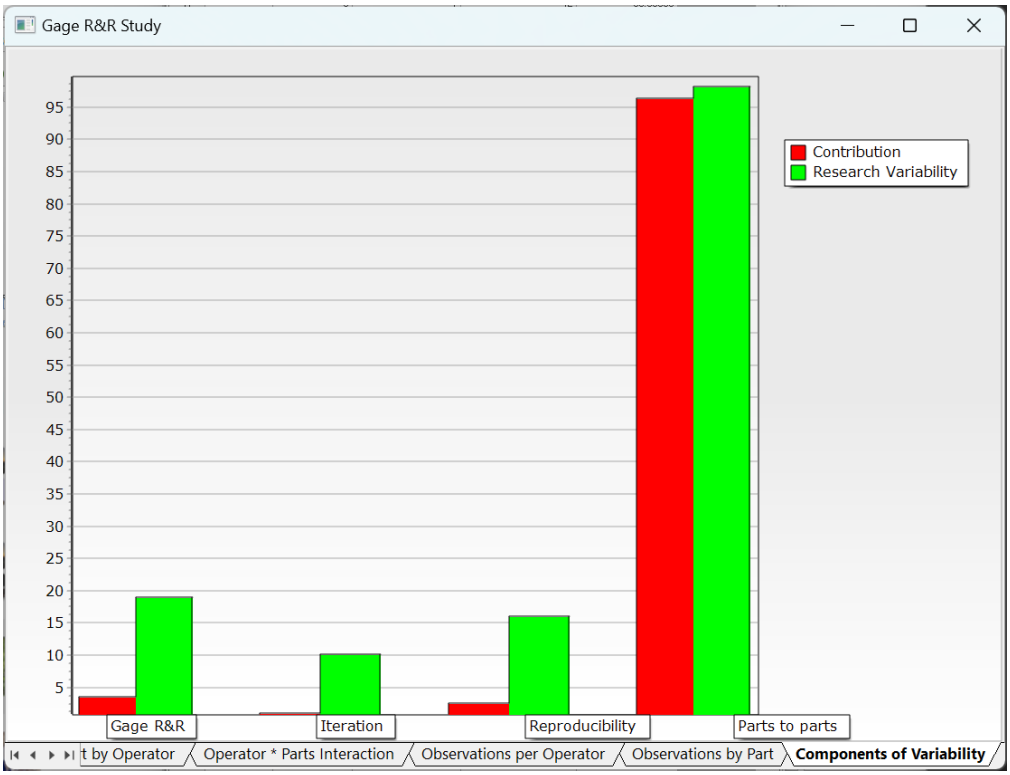
  Displays observations per part only, with the solid line representing the mean.



-

- **Components of Variability**

  Displays where the variation in observed values originates.

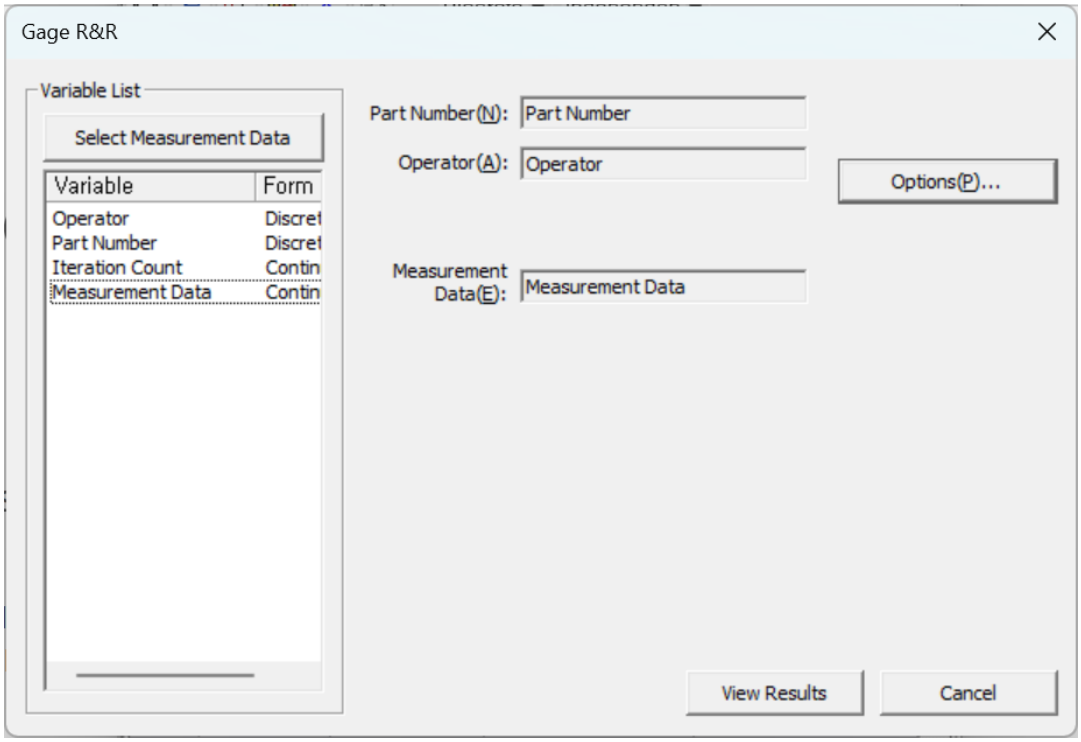


4.3.11.4 Gage R&R Study (Nested Design)

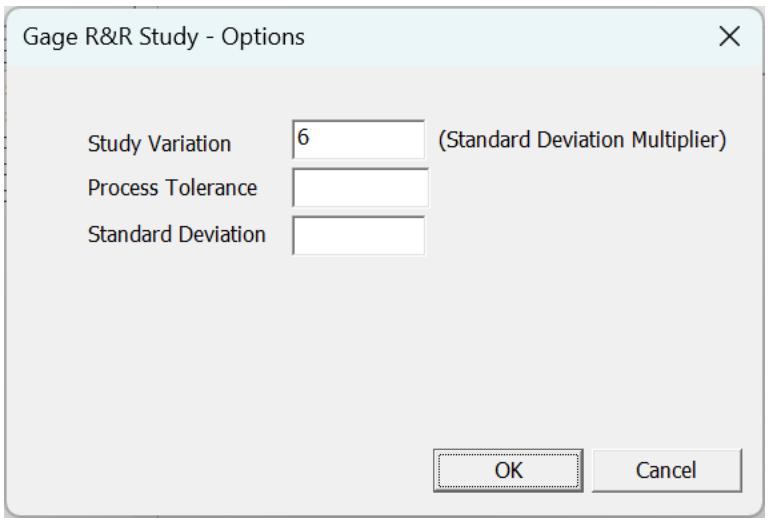
Gage R&R Study (Nested Design) is a technique which each part is measured by only one operator. It is commonly used in destructive testing and in systems where once one operator measures a part, no other operator can measure the same part. It is essential to assume that parts within each batch are nearly identical. If this assumption is violated, operators will measure different parts, making it unclear whether the variation is due to part differences or issues in the measurement system. In nested designs, total variation is divided into part-to-part, reproducibility, and repeatability, allowing for accurate identification of measurement variability sources.

How to run

[Analyze] – [Gage R&R] – [Gage R&R Study (Nested Design)]



- **Part Number**: Select the parts (Must be discrete)
- **Operator**: Select the one who takes measures (Must be discrete)
- **Measurement Data**: Select the measured data (Must be continuous)

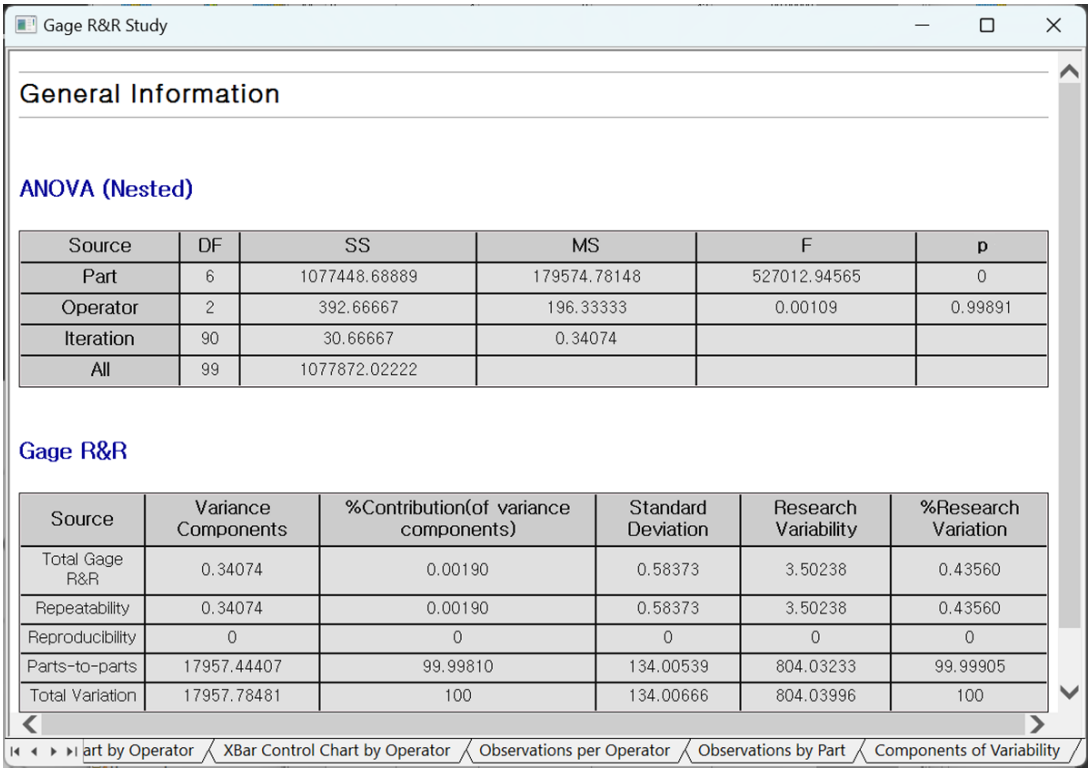


- **Study Variation**: Enter the coefficient to obtain the Study Variation.
- **Process Tolerance**: Enter an empirically known tolerance to calculate %tolerance.
- **Standard Deviation**: Enter an empirically known Historical Standard to calculate %process.

Results

- **General Information**

  Shows the results of ANOVA and Gage R&R analysis results.



Gage R&R Study

General Information

ANOVA (Nested)

Source	DF	SS	MS	F	p
Part	6	1077448.68889	179574.78148	527012.94565	0
Operator	2	392.66667	196.33333	0.00109	0.99891
Iteration	90	30.66667	0.34074		
All	99	1077872.02222			

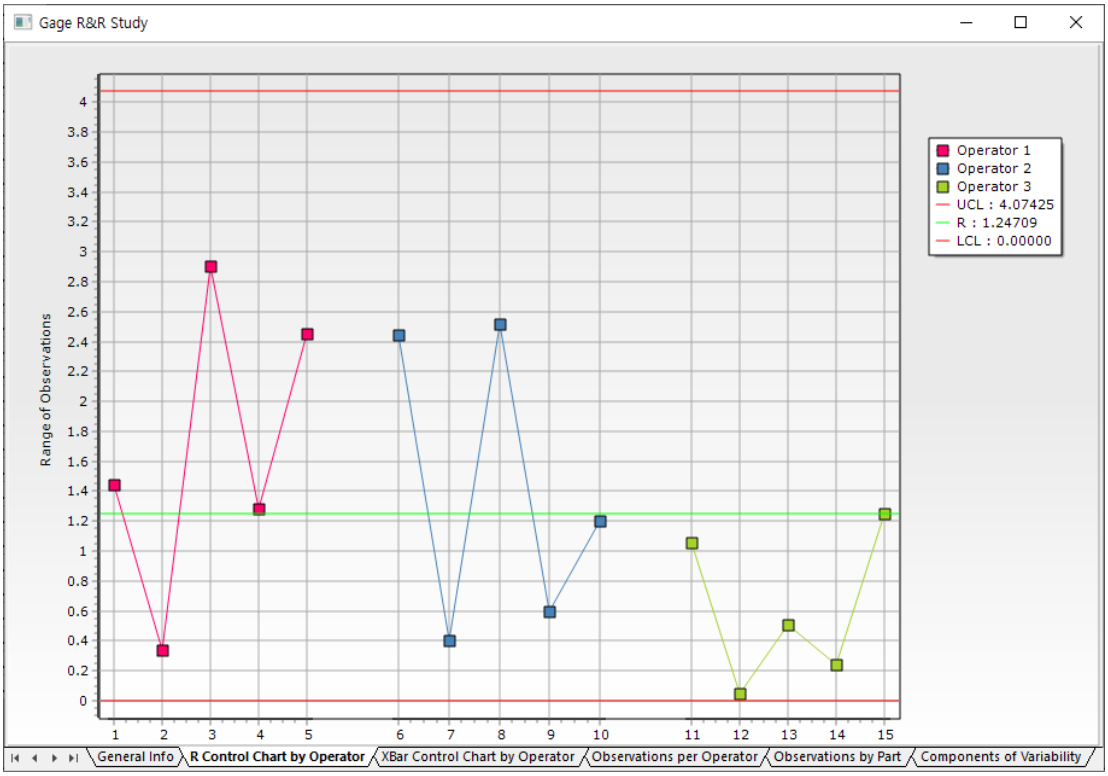
Gage R&R

Source	Variance Components	%Contribution(of variance components)	Standard Deviation	Research Variability	%Research Variation
Total Gage R&R	0.34074	0.00190	0.58373	3.50238	0.43560
Repeatability	0.34074	0.00190	0.58373	3.50238	0.43560
Reproducibility	0	0	0	0	0
Parts-to-parts	17957.44407	99.99810	134.00539	804.03233	99.99905
Total Variation	17957.78481	100	134.00666	804.03996	100

art by Operator | XBar Control Chart by Operator | Observations per Operator | Observations by Part | Components of Variability

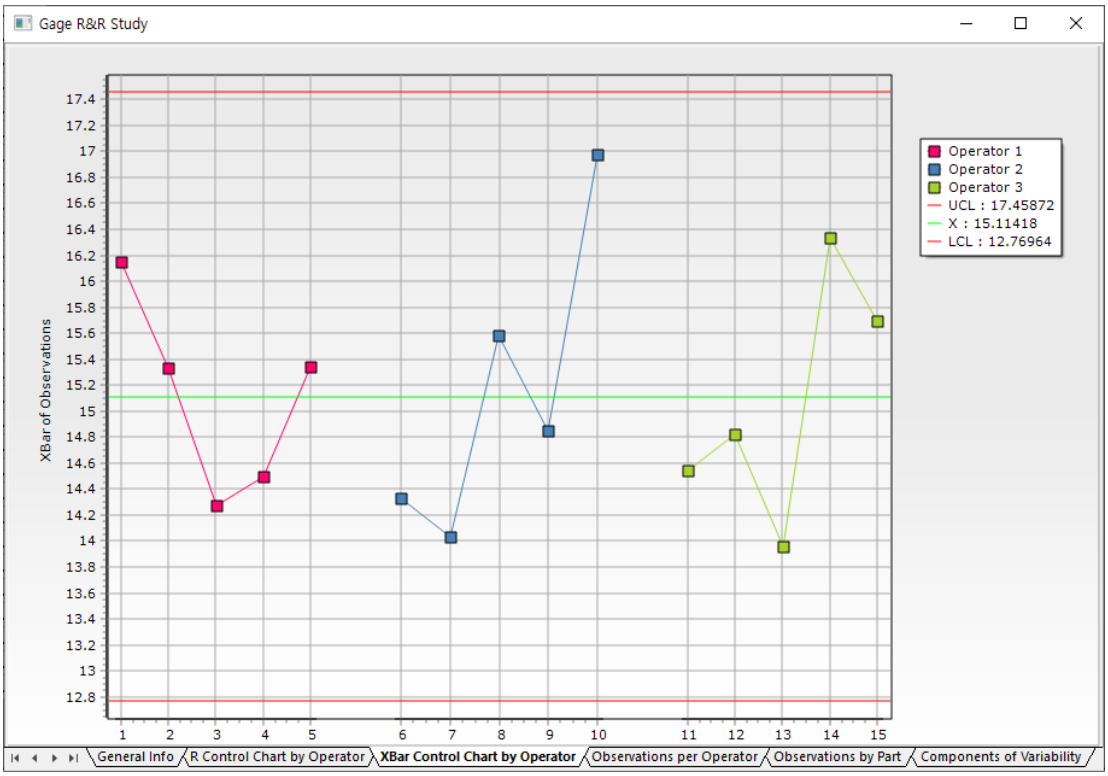
- **R Control Chart by Operator**

  Each point displays the difference in the smallest and the biggest measurement of the same part by each operator.



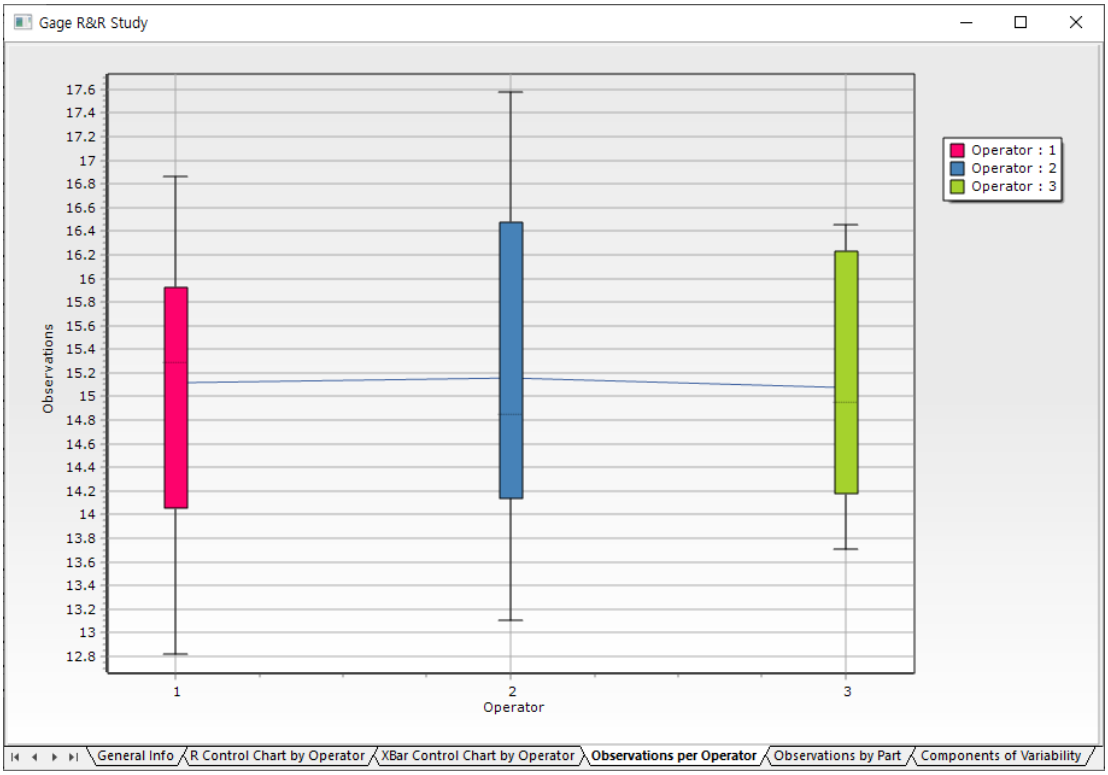
- **X-Bar Control Chart by Operator**

  Each point displays the average measurement of the same part by each operator.



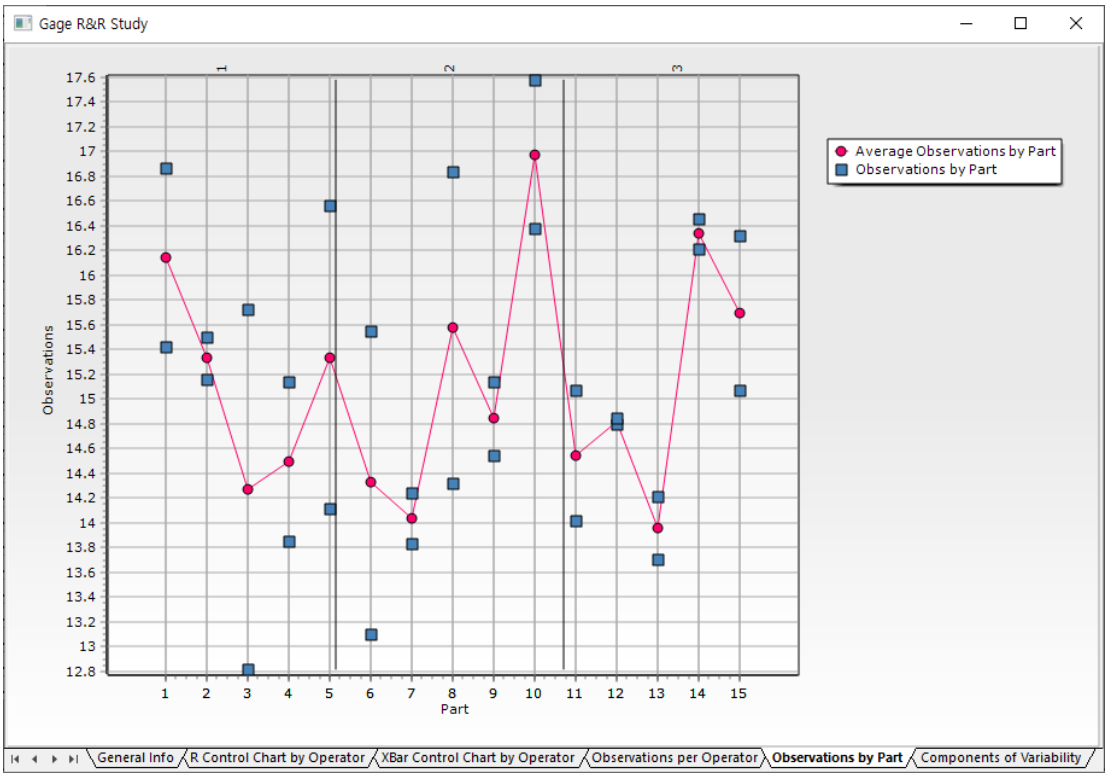
- **Observations per Operator**

  Displays box plots of all measurements for each operator, with the solid line representing the mean.



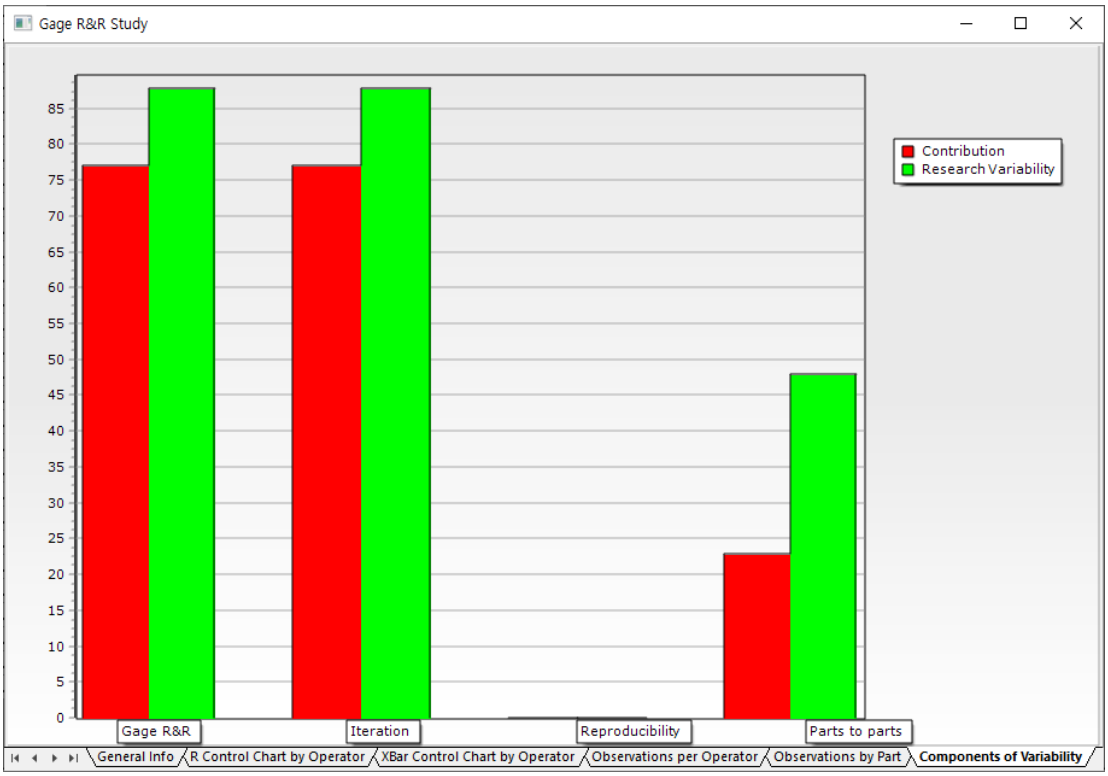
- **Observations by Part**

  Displays observations per part only, with the solid line representing the mean.



- **Components of Variability**

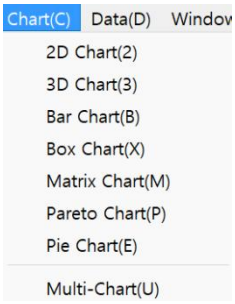
  Displays where the variation in observed values originates.



4.4 Chart

Data Browser provides various charting functions.

Basic Charts



2D chart, 3D chart, Bar chart, Box chart, Matrix chart, Pareto chart, Pie chart and Multi-chart functions.

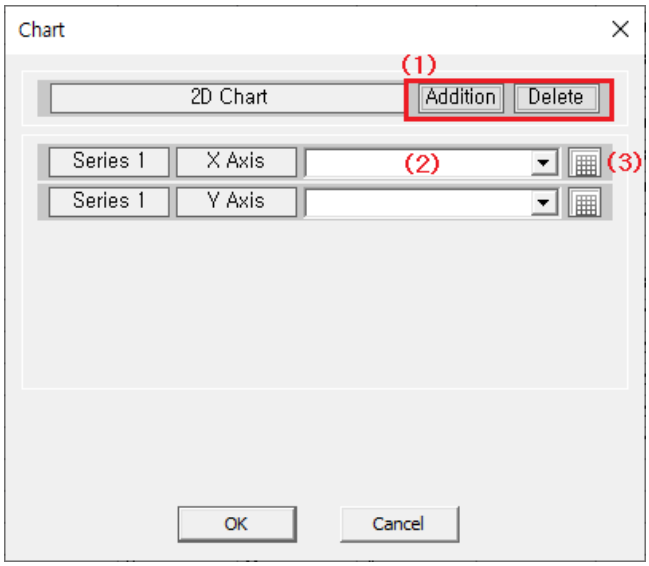
4.4.1 Basic Charts

How to run

[Chart]

How to choose chart options

The following example is a two-dimensional chart.



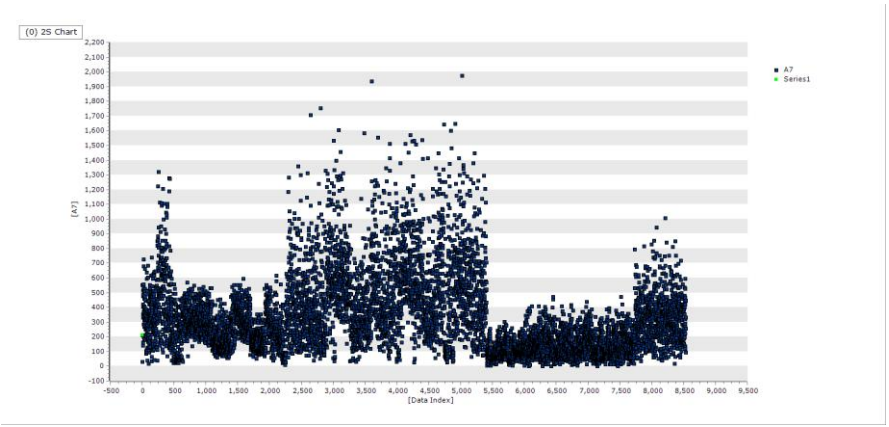
1: Add/delete. Add more chart or delete chart.

2: Select variables for chart

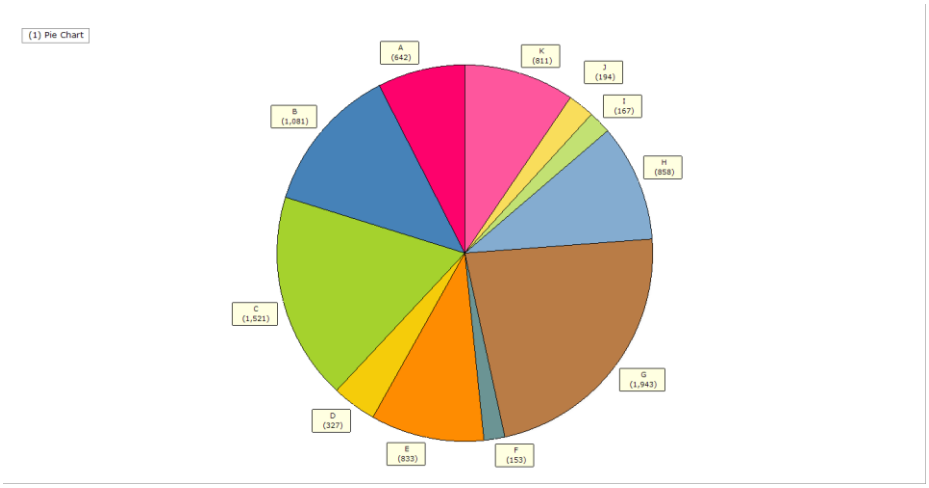
3: You may select a variable field in data browser instead selecting a variable.

Results

- **2D chart**



- **Pie chart**

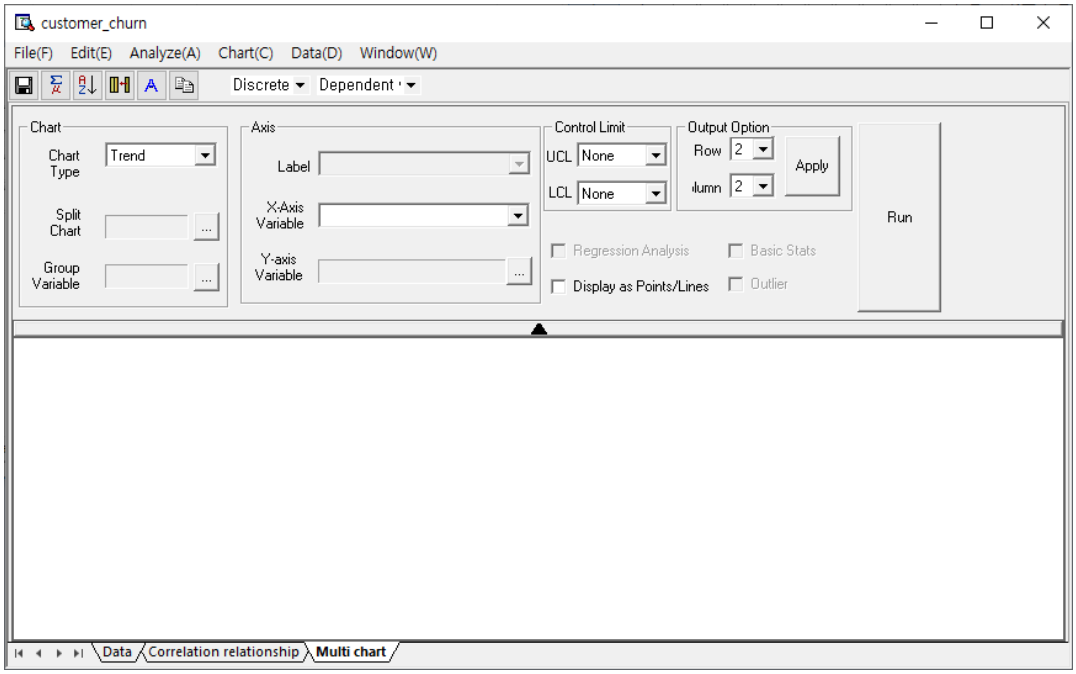


4.4.2 Multi chart

Multi-chart is an unique chart of ECMiner™ which makes it easy to see the characteristics of multiple variables.

How to run

[Analyze] – [Chart] - [Multi-chart]



Chart

- **Chart Type**

  Trend, Data, Box Plot, Distribution, and Correlation charts.

  Data: Displays data as a scatter plot.

  Box Plot: Visualizes data distribution.

  Distribution: Shows normal distribution of variables.

  Correlation: Highlights relationships between variables.

- **Split Chart:**

  Select a group variable. Items with the same variable value (e.g., 1 or -1) are drawn separately.

- **Group Variable:**

  Select a group variable. Items with the same value are shown as dots of the same color.

- **X Axis**

  Label: Select a discrete variable as the X-axis in the chart.

  Order: This is a variable that determines in what order the variables in the label will be displayed on the X-axis.

- **Y Axis**

  Y1: Select a variable on the left Y-axis. Multiple Y-axis selections are possible.

  Y2: Select a variable on the right Y-axis.

  X, **Y**: Select Correlation as the chart type, decide which variable to use as the X-axis and which variable to use as the Y-axis.

Check Box Options

- **Regression Analysis**

  Options for regression function.

- **Display as Points/Lines**

  Options in the plot for connecting points with lines.

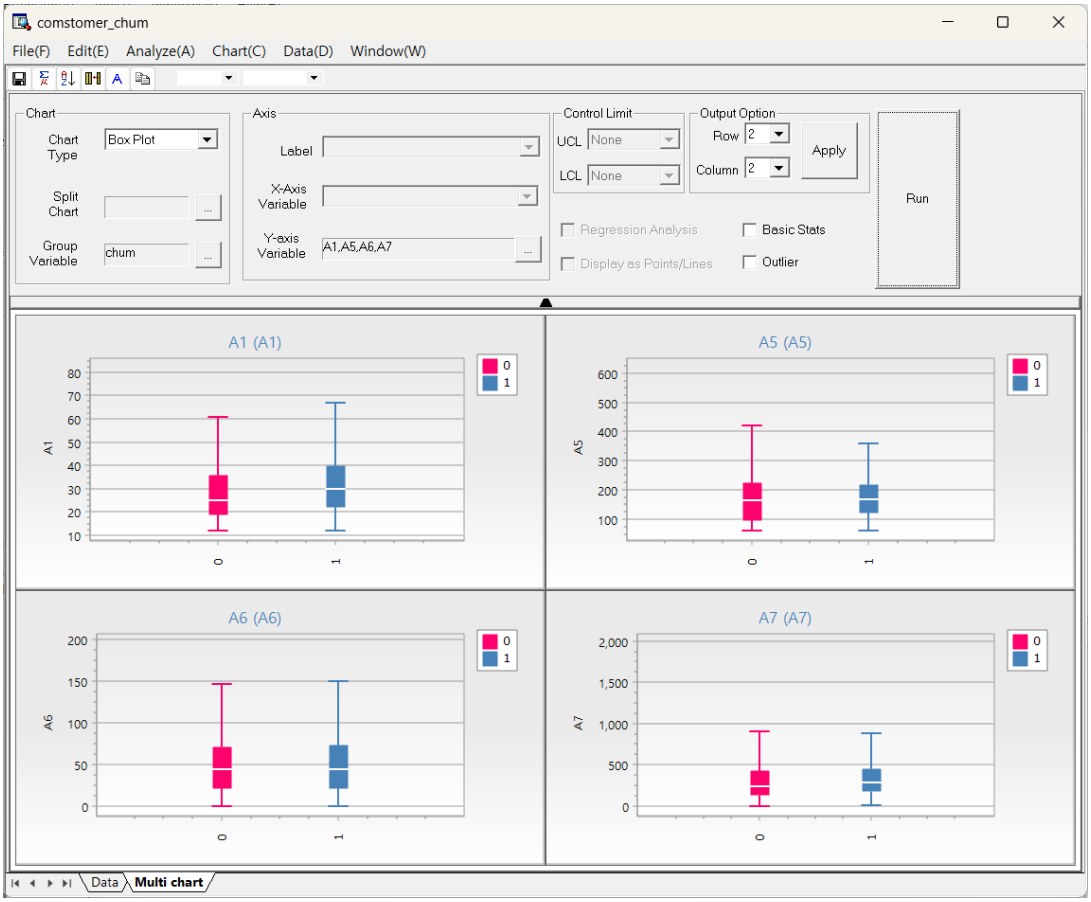
- **Basic Statistics**

  Options for descriptive statistics with Median, Standard Deviation, Average, Maximum, Minimum, and Range.

- **Outlier**

  Options for displaying outliers.

- **The following is the Box Chart screen when 'Basic Statistics' and 'Outliers' are selected.**



- **Other Options**

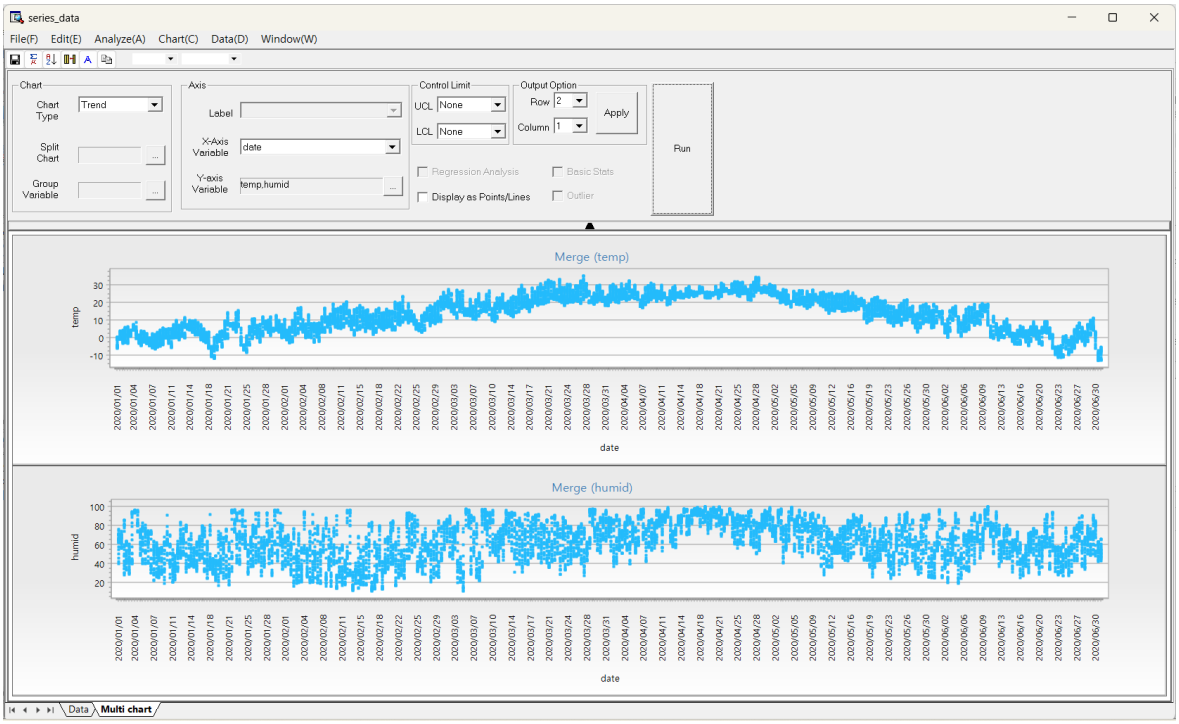
  Output Option (Row): Set how many charts to show horizontally.

  Output Option (Column): Set how many charts to show vertically.

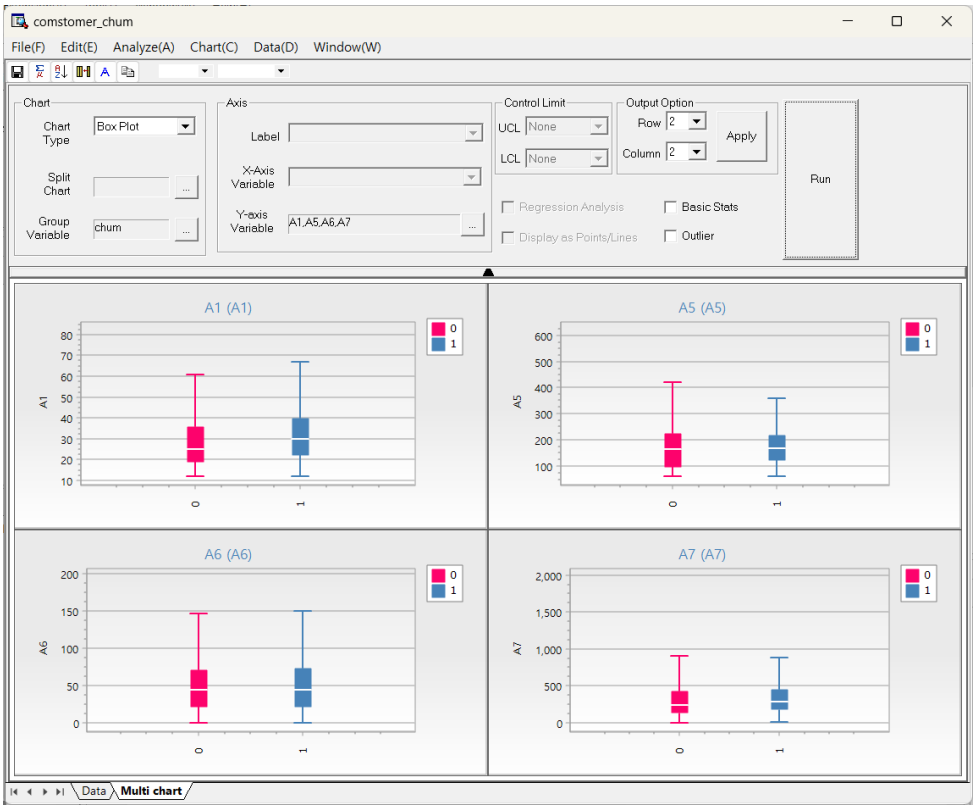
  Control limit: Options for the basis for displaying the control line.

Example

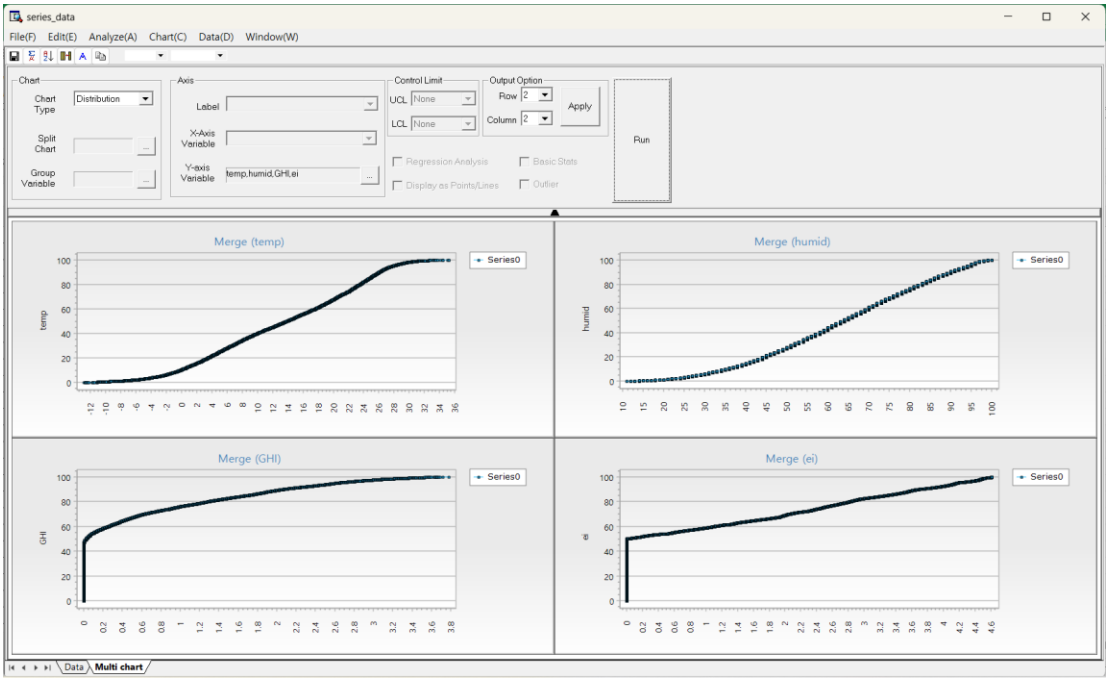
- **Trend chart type.**



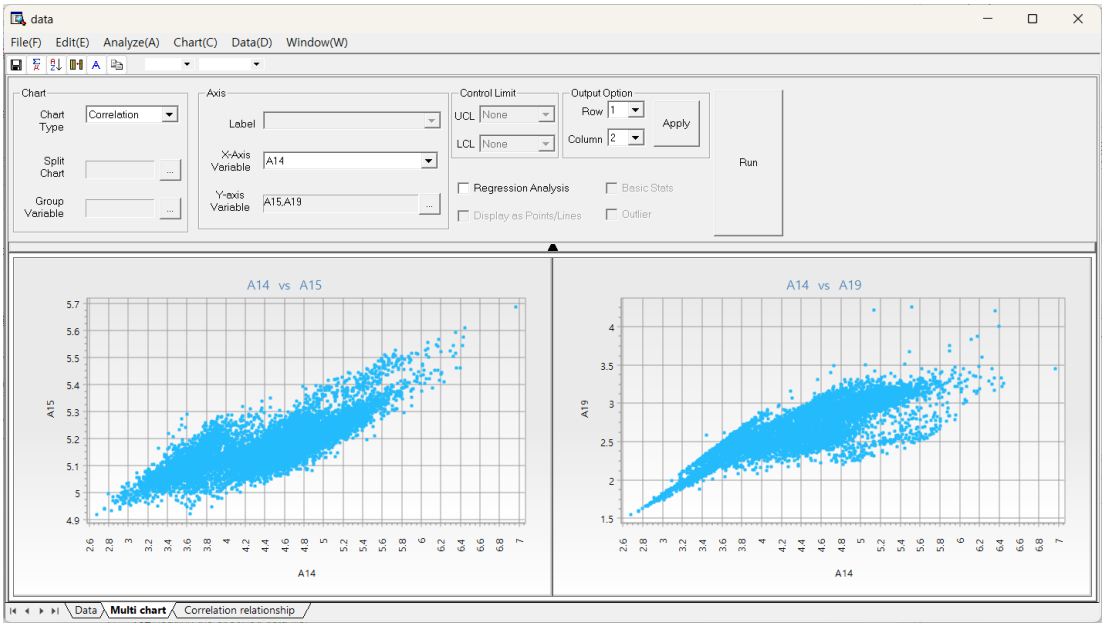
- **Box plot chart type.**



- **Distribution chart type.**



- **Correlation chart type.**



4.5 Data

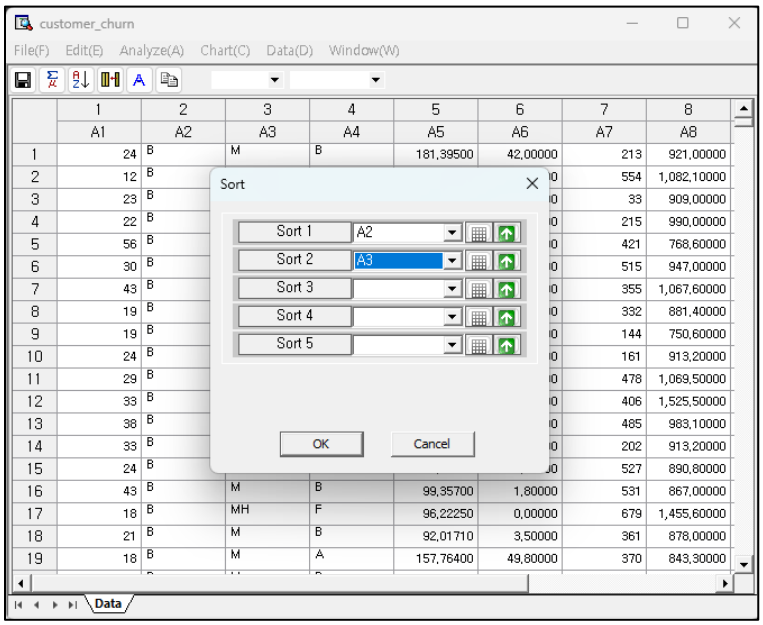
4.5.1 Data Sorting

The Data Browser supports multiple column sorting. This function works the same as the Sort Node in the Preprocessing Nodes.

How to run

[Data] - [Sort]

Sort the data in ascending or descending order by clicking the direction icon. To sort by multiple fields, select additional fields for multi-level sorting.

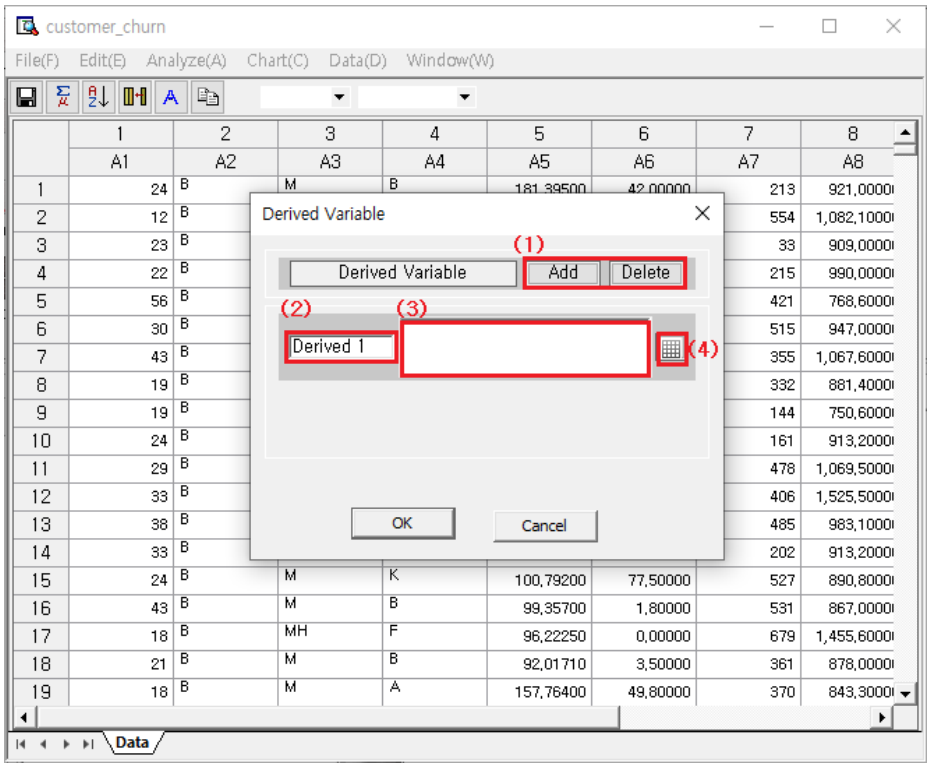


4.5.2 Derived Variables

The Data Browser supports creating derived variables. This function works the same as the Derived Column Node in the Preprocessing Nodes.

How to run

[Data] - [Derived Variables]



- **(1) Management of derived variables**

  Add for creating a new derived variable

  Delete for removing a derived variable

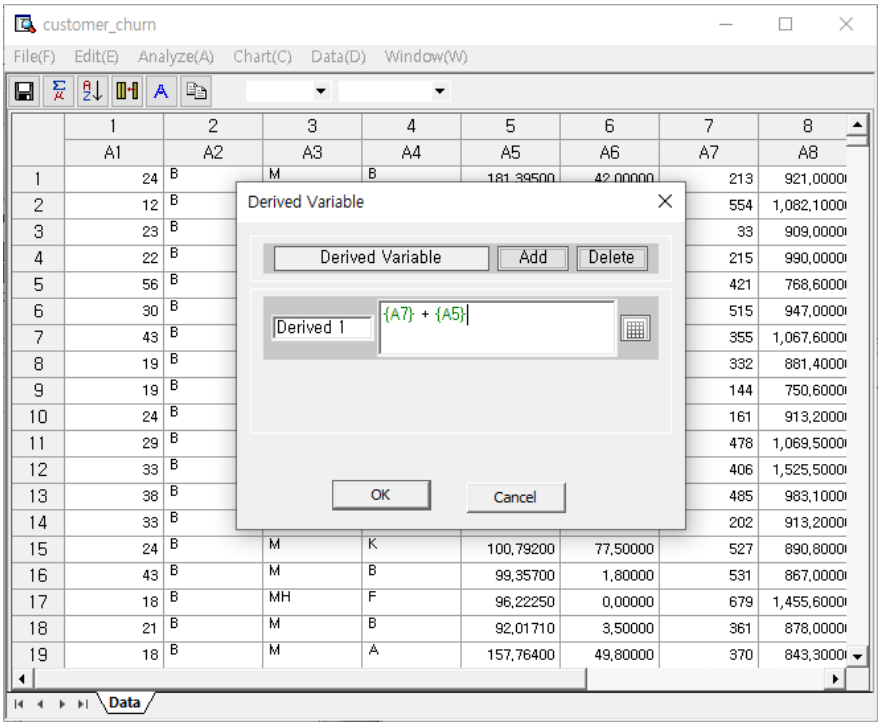
  If only one list remains, pressing the delete button will not remove it.

- **(2) Name of a derived variable**

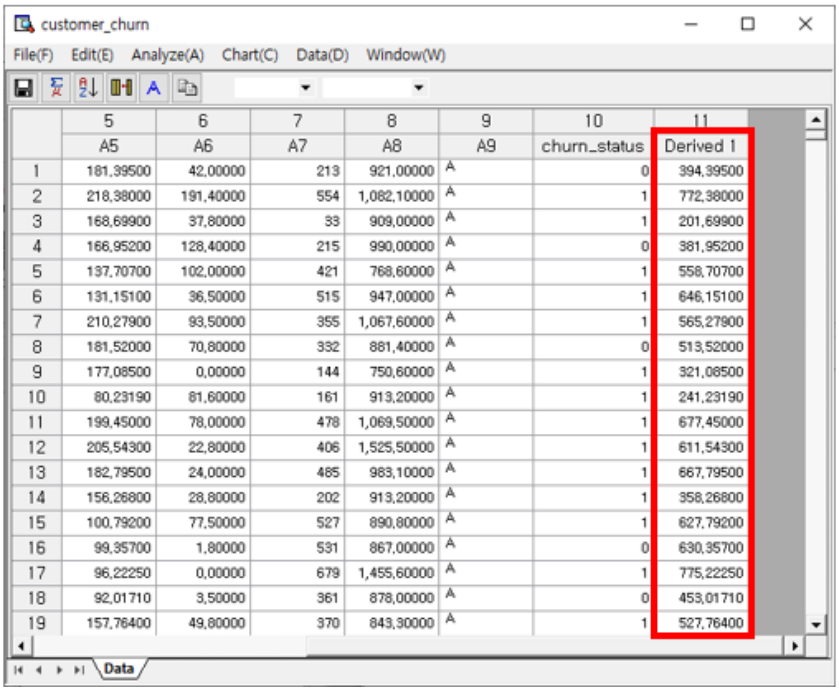
  Give a name of the derived variable

- **(3) Expression Editor**

  Use the expression editor to define how to calculate the derived variable



Results



	5	6	7	8	9	10	11
	A5	A6	A7	A8	A9	churn_status	Derived 1
1	181,39500	42,00000	213	921,00000	A	0	394,39500
2	218,38000	191,40000	554	1,082,10000	A	1	772,38000
3	168,69900	37,80000	33	909,00000	A	1	201,69900
4	166,95200	128,40000	215	990,00000	A	0	381,95200
5	137,70700	102,00000	421	768,60000	A	1	558,70700
6	131,15100	36,50000	515	947,00000	A	1	646,15100
7	210,27900	93,50000	355	1,067,60000	A	1	565,27900
8	181,52000	70,80000	332	881,40000	A	0	513,52000
9	177,08500	0,00000	144	750,60000	A	1	321,08500
10	80,23190	81,60000	161	913,20000	A	1	241,23190
11	199,45000	78,00000	478	1,069,50000	A	1	677,45000
12	205,54300	22,80000	406	1,525,50000	A	1	611,54300
13	182,79500	24,00000	485	983,10000	A	1	667,79500
14	156,26800	28,80000	202	913,20000	A	1	358,26800
15	100,79200	77,50000	527	890,80000	A	1	627,79200
16	99,35700	1,80000	531	867,00000	A	0	630,35700
17	96,22250	0,00000	679	1,455,60000	A	1	775,22250
18	92,01710	3,50000	361	878,00000	A	0	453,01710
19	157,76400	49,80000	370	843,30000	A	1	527,76400

4.5.3 Apply

The Data Browser supports creation of streams onto the Workspace. The preprocessing functions that are applied in the Data Browser can be created into a stream with the Apply function.

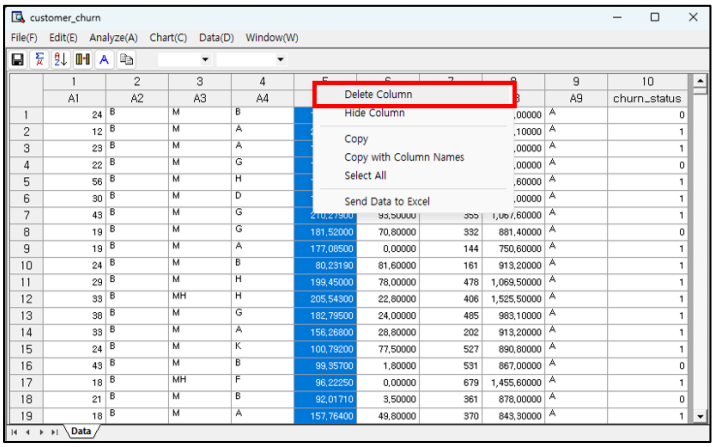
How to run

[Data] - [Apply]

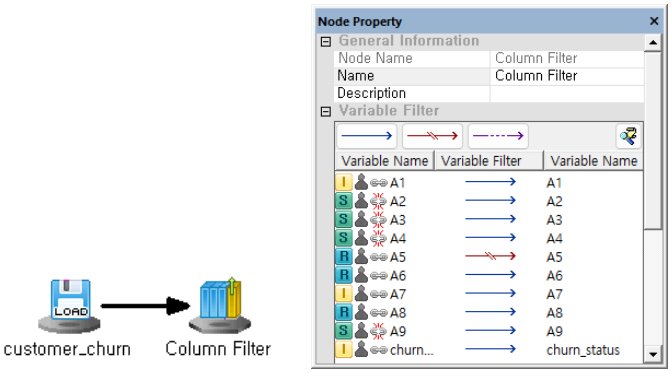
Results

Examples

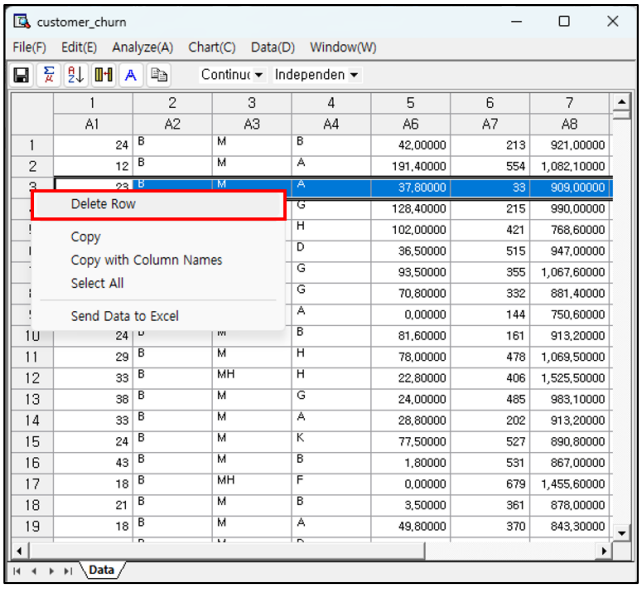
- **Delete column**



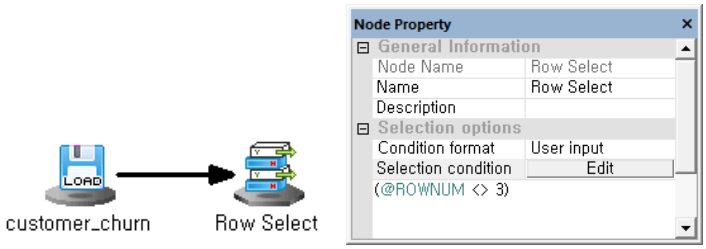
Press right click on the column that you want to delete and click 'Delete Column' in the Data Browser. Then run the **Apply** function to create a stream that applies the preprocessing methods. In the Workspace, the **File Reader Node** is automatically connected to the **Column Filter Node** that deletes the selected column.



- **Delete row**

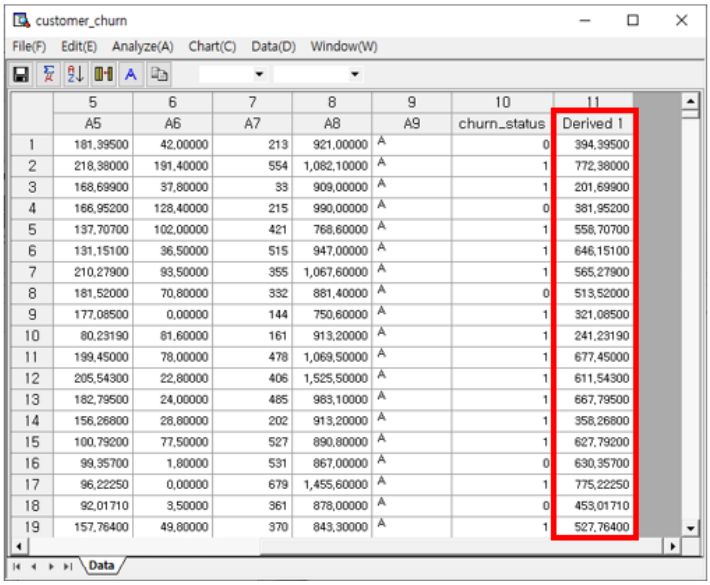


Press right click on the row that you want to delete and click 'Delete Row' in the Data Browser. Then run the **Apply** function to create a stream that applies the preprocessing methods. In the Workspace, the **File Reader Node** is automatically connected to the **Row Select Node** that deletes the selected column.



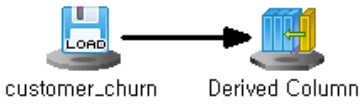
NOTE (@ROWNUM <> 3) is a conditional statement that means "a row whose row index is not 3".

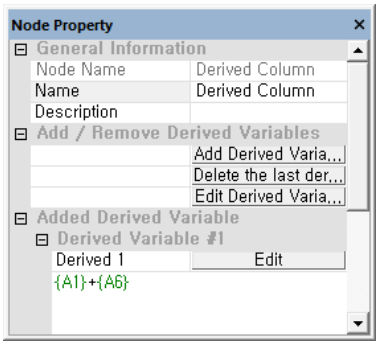
- **Derived Variables**



	5	6	7	8	9	10	11
	A5	A6	A7	A8	A9	churn_status	Derived 1
1	181,39500	42,00000	213	921,00000	A	0	394,39500
2	218,38000	191,40000	554	1,082,10000	A	1	772,38000
3	168,69900	37,80000	33	909,00000	A	1	201,69900
4	166,95200	128,40000	215	990,00000	A	0	381,95200
5	137,70700	102,00000	421	768,60000	A	1	558,70700
6	131,15100	36,50000	515	947,00000	A	1	646,15100
7	210,27900	93,50000	355	1,067,60000	A	1	565,27900
8	181,52000	70,80000	332	881,40000	A	0	513,52000
9	177,08500	0,00000	144	750,60000	A	1	321,08500
10	80,23190	81,60000	161	913,20000	A	1	241,23190
11	199,45000	78,00000	478	1,069,50000	A	1	677,45000
12	205,54300	22,80000	406	1,525,50000	A	1	611,54300
13	182,79500	24,00000	485	983,10000	A	1	667,79500
14	156,26800	28,80000	202	913,20000	A	1	358,26800
15	100,79200	77,50000	527	890,80000	A	1	627,79200
16	99,35700	1,80000	531	867,00000	A	0	630,35700
17	96,22250	0,00000	679	1,455,60000	A	1	775,22250
18	92,01710	3,50000	361	878,00000	A	0	453,01710
19	157,76400	49,80000	370	843,30000	A	1	527,76400

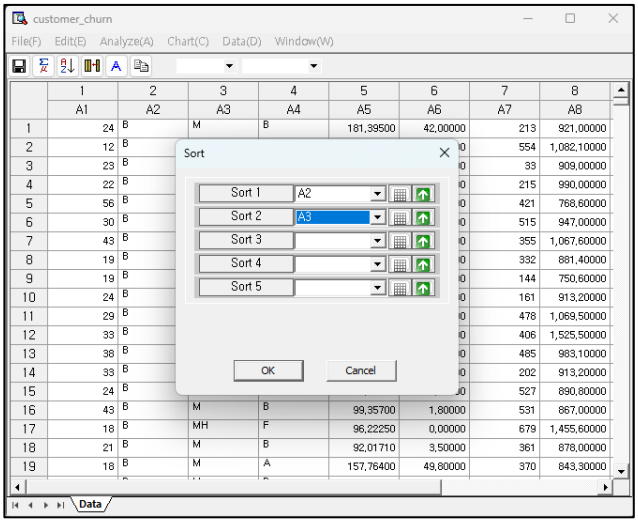
After the Derived Variable function in the Data Browser is applied, run the **Apply** function to create a stream. In the Workspace, the **File Reader Node** is automatically connected to the **Derived Column Node** that creates a new variable with the defined conditional statement.



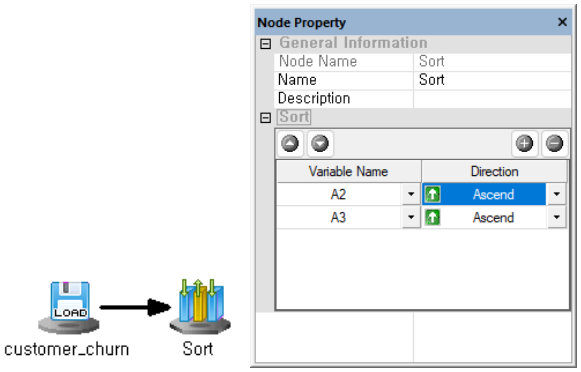


NOTE (*{A1} + {A6}*) is a conditional statement that means "add the values of A1 and A6 variables".

- **Sort**



After the Sort function in the Data Browser is applied, run the **Apply** function to create a stream. In the Workspace, the **File Reader Node** is automatically connected to the **Align Node** that sorts the selected variable.



- **Apply multiple preprocessing**

Preprocessing results can be applied one by one as in the example above, but multiple preprocessing results can also be applied at once.

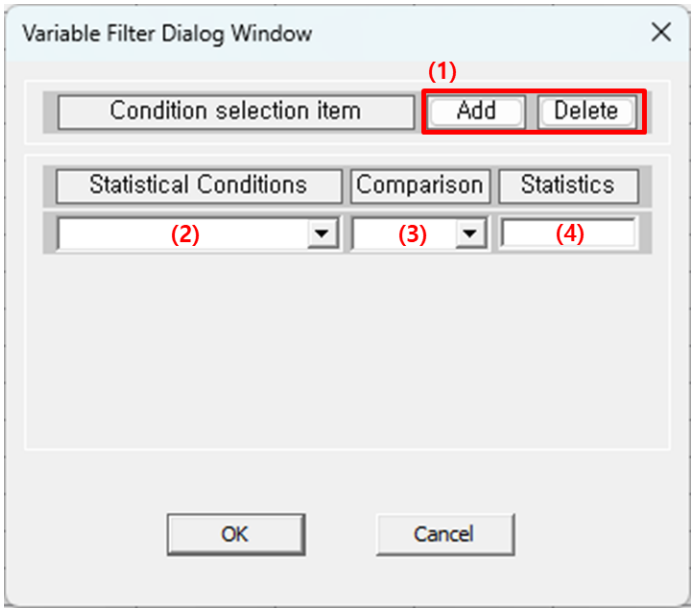
Below is an example of the created stream after applying multiple preprocessing

4.5.4 Filter

The Data Browser supports Filtering. This function works similar to the Derived Column Node in the Preprocessing Nodes, but in the data browser the filter function applies to all variables.

How to run

[Data] – [Filter]



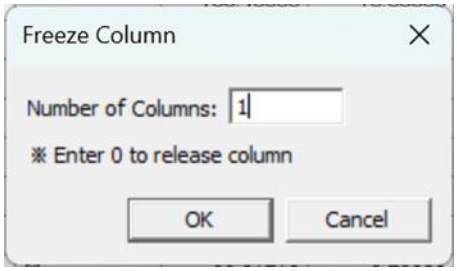
- 1. "Add" to add a condition. "Delete" to remove the last one
- 2. Select statistical conditions, such as standard deviation.
- 3. Choose comparison operators >, =, <, etc.
- 4. Specify the value for comparison.

4.5.5 Freeze Columns

The more data columns you have, the more likely you'll need to scroll horizontally to view variables. When comparing distant columns, it's difficult to do so visually. With 'Freeze Coumns,' the specified column is pinned, allowing the remaining columns to be visible when scrolling.

How to run

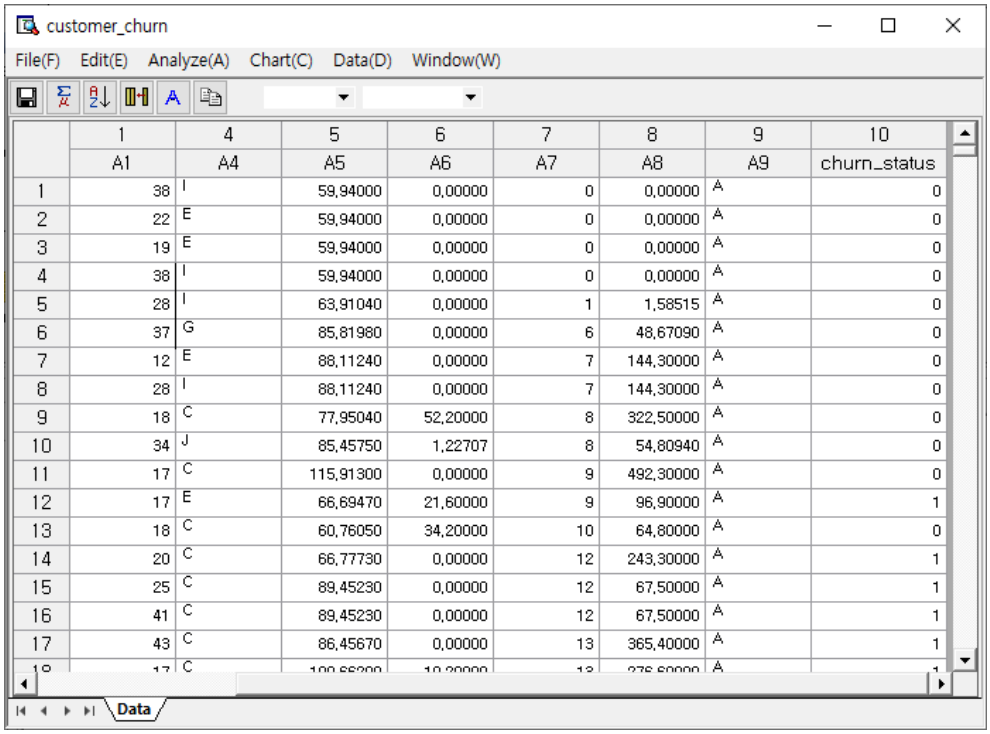
[Data] - [Freeze Columns]



Number of Columns: Select how many columns to freeze from the first column.

Results

If you freeze 1 column, you can see that the first columns ‘A1’ is fixed to the left even when you move the horizontal scroll bar to the right.



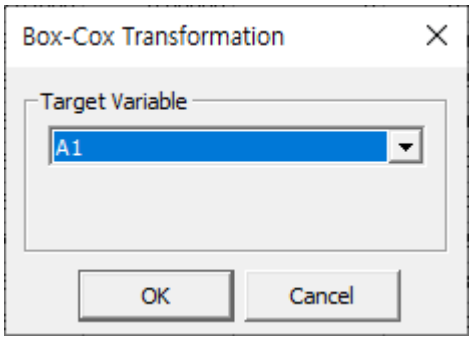
	1	4	5	6	7	8	9	10
	A1	A4	A5	A6	A7	A8	A9	churn_status
1	38	I	59,94000	0,00000	0	0,00000	A	0
2	22	E	59,94000	0,00000	0	0,00000	A	0
3	19	E	59,94000	0,00000	0	0,00000	A	0
4	38	I	59,94000	0,00000	0	0,00000	A	0
5	28	I	63,91040	0,00000	1	1,58515	A	0
6	37	G	85,81980	0,00000	6	48,67090	A	0
7	12	E	88,11240	0,00000	7	144,30000	A	0
8	28	I	88,11240	0,00000	7	144,30000	A	0
9	18	C	77,95040	52,20000	8	322,50000	A	0
10	34	J	85,45750	1,22707	8	54,80940	A	0
11	17	C	115,91300	0,00000	9	492,30000	A	0
12	17	E	66,69470	21,60000	9	96,90000	A	1
13	18	C	60,76050	34,20000	10	64,80000	A	0
14	20	C	66,77730	0,00000	12	243,30000	A	1
15	25	C	89,45230	0,00000	12	67,50000	A	1
16	41	C	89,45230	0,00000	12	67,50000	A	1
17	43	C	86,45670	0,00000	13	365,40000	A	1

4.5.6 Box-Cox Transformation

To apply a control chart, the data must follow a normal distribution, which is often not the case. The Box-Cox Transformation provides a method to convert non-normal data into a normal distribution.

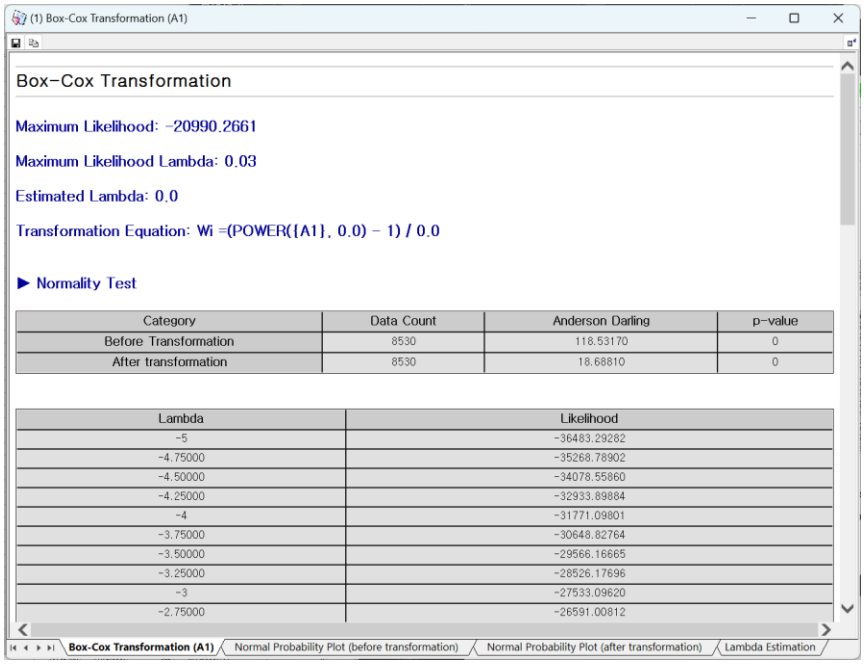
How to run

[Data] – [Box-Cox Transformation]



Results

It provides the estimated lambda to define the transformation formula and offers normality test statistics before and after transformation, showing the improvement in normality. The conversion formula can be applied for a new variable using the Derived Variable Node.

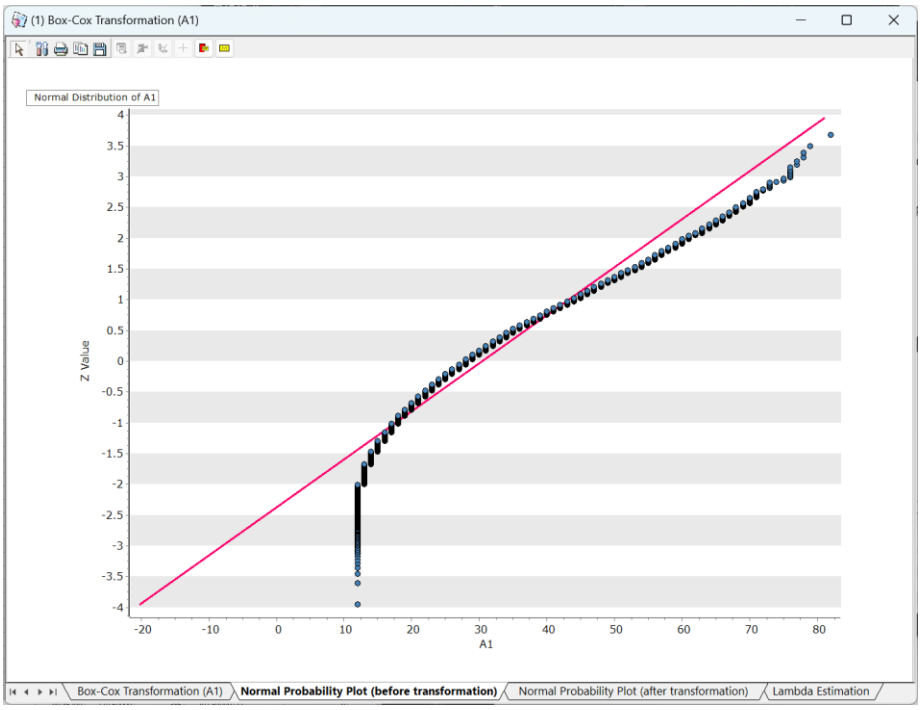


Category	Data Count	Anderson Darling	p-value
Before Transformation	8530	118.53170	0
After transformation	8530	18.68810	0

Lambda	Likelihood
-5	-36483.29282
-4.75000	-35268.78902
-4.50000	-34078.55860
-4.25000	-32933.89884
-4	-31771.09801
-3.75000	-30648.82764
-3.50000	-29566.16665
-3.25000	-28526.17696
-3	-27533.09620
-2.75000	-26591.00812

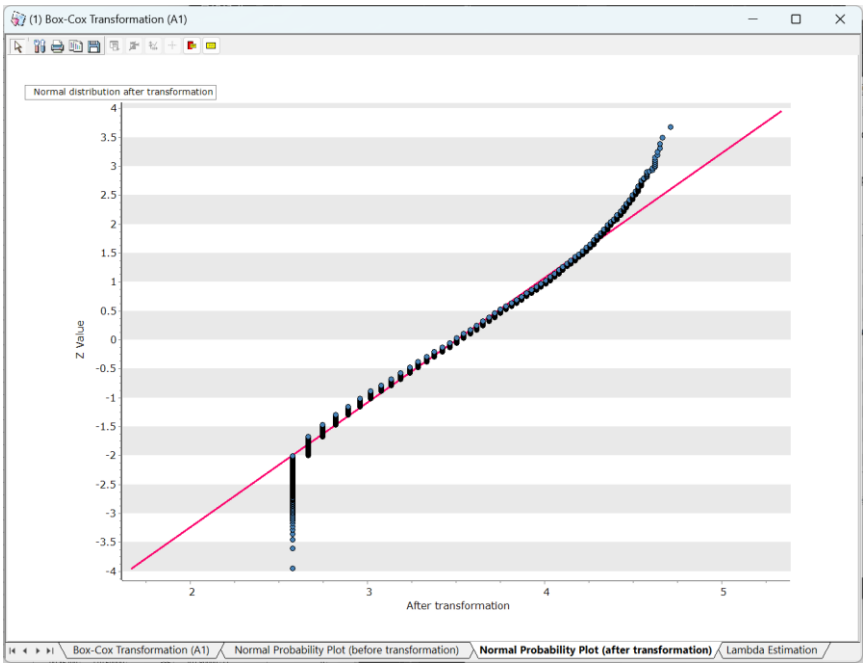
- **Normal Probability Plot (before transformation)**

  This graph shows how closely the data follows a normal distribution before transformation. If a lot of data is distributed around the red line, the data can be said to follow normality. The more it deviates from the red line, the more it violates the assumption of normality.



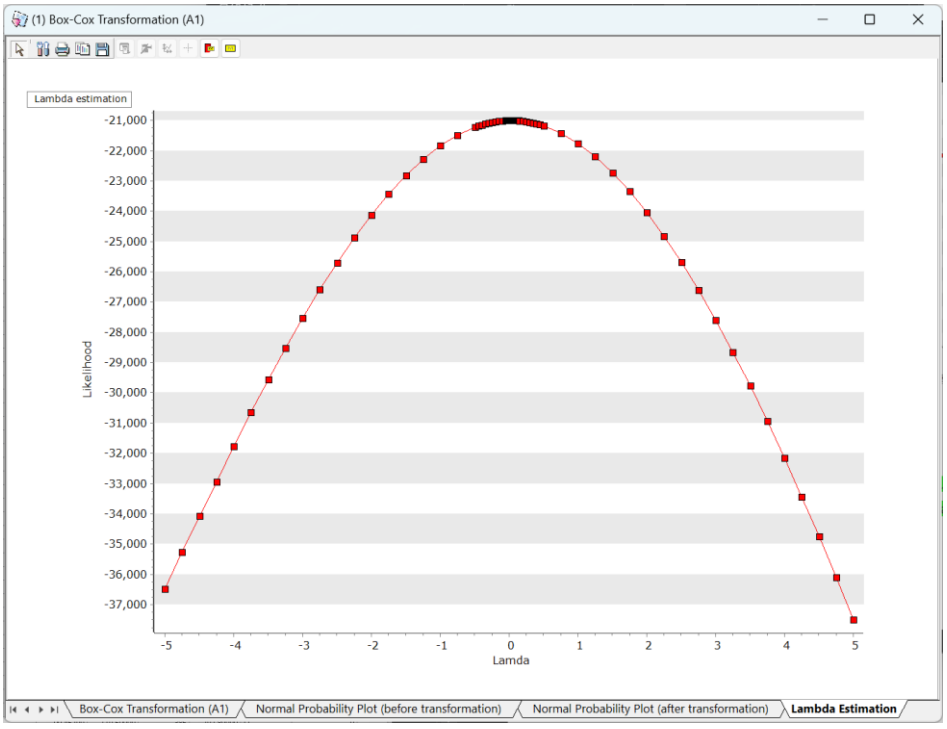
- **Normal Probability Plot (after Transformation)**

  This graph shows how well the data transformed by the estimated transformation formula follows the normal distribution. If a lot of data is distributed around the red line, the data can be said to follow normality. The more it deviates from the red line, the more it violates the assumption of normality.



- **Lambda Estimation**

  The transformation formula is determined using the lambda with the maximum likelihood as an estimate, and a graph showing the change in likelihood depending on the lambda is provided.

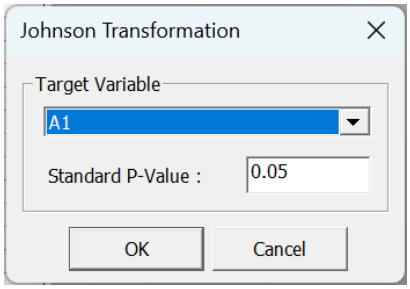


4.5.7 Johnson Transformation

The Johnson Transformation offers a formula to convert non-normal data into a normal distribution, allowing it to be used in control charts.

How to run

[Data] – [Johnson Transformation]



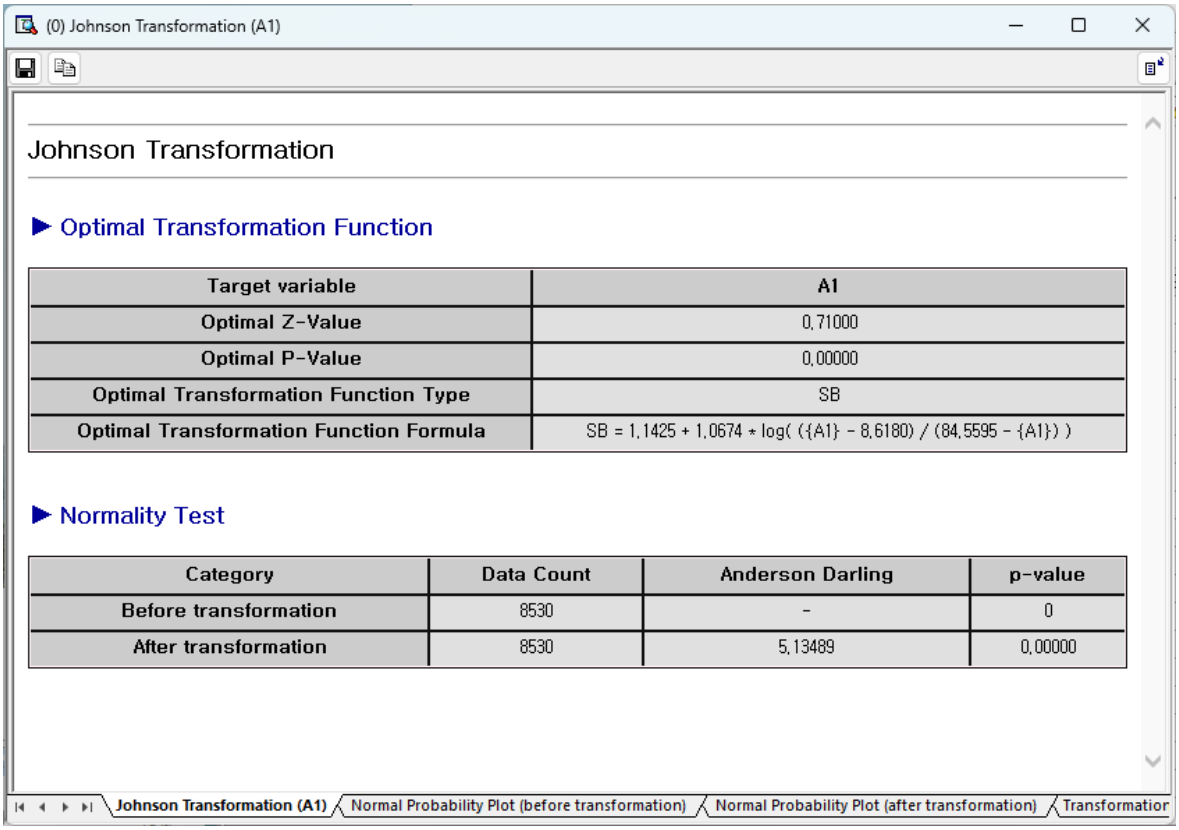
Results

- **Optimal Transformation Function**

  Provides the estimated transformation function that best fits a normal distribution. This function can be applied as a transformed variable using the Derived Variable Node.

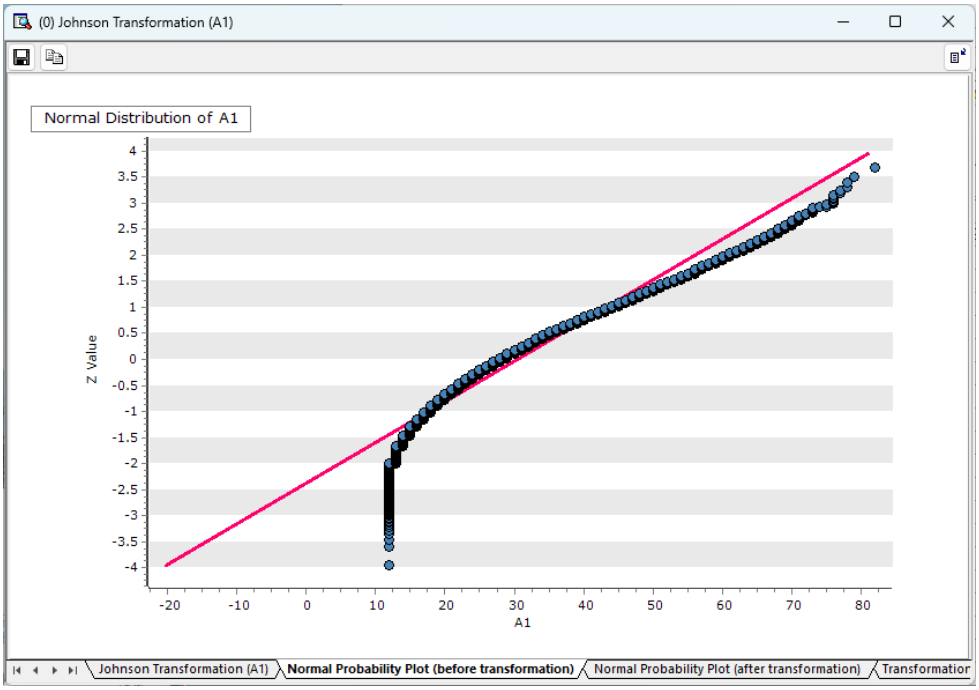
- **Normality Test**

  Anderson-Darling statistic and p-value to evaluate how the variable follows a normal distribution. A p-value greater than 0.05 indicates normality.



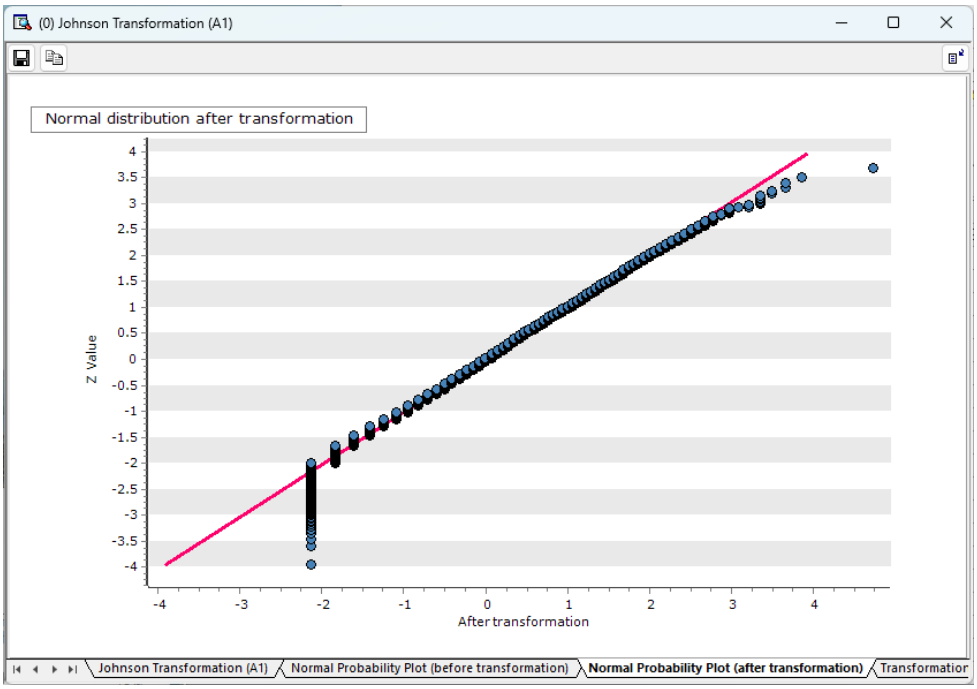
- **Normal Probability Plot (before transformation)**

  This graph shows how closely the data follows a normal distribution before transformation. If a lot of data is distributed around the red line, the data can be said to follow normality. The more it deviates from the red line, the more it violates the assumption of normality.



- **Normal Probability Plot (after transformation)**

  This graph shows how the transformed data follows the normal distribution. The more it deviates from the red line, the more it violates the assumption of normality.



- **Transformation Estimate**

Provides a graph that allows you to understand the extent to which transformation data that varies depending on the Z value follows a normal distribution (p-value). The larger the p-value, the more it can be said to satisfy normal distribution.