

Predictive Statistics with Toy Company Sales Dataset/ SAS Studio

Didem B. Aykurt

Colorado State University Global

MIS543: Enterprise Performance Management

Dr. John Marlowe

October 8, 2023

Predictive Statistics in SAS Studio

Predictive statistics is a branch that uses statistical models to predict future events or outcomes based on historical data. Predictive statistics are used in various applications, including finance, marketing, healthcare, and more. Predictive statistics involves using statistical models to identify patterns and trends in historical data and use them to make predictions about future events or outcomes. These models can forecast future sales, customer behavior, and other business metrics. Many predictive models exist, including decision trees, time series, regression analysis, and neural networks. These models use different statistical techniques to analyze data and predict future events or outcomes. Predictive statistics are a powerful tool that can help organizations make better decisions and improve business outcomes. Organizations can make more informed decisions about the future and optimize operations by analyzing historical data and identifying patterns and trends. Some common applications of predictive statistics include:

Forecasting: Predictive analytics can forecast future sales, customer behavior, and other business metrics. This can help organizations make better decisions and optimize their operations.

Marketing: Predictive analytics can identify customer segments and predict which customers will most likely purchase a product or service. This can help organizations develop more effective marketing strategies and improve customer engagement.

Fraud detection: Predictive analytics can detect fraudulent activity by analyzing patterns and trends in data. This can help organizations reduce the risk of fraud and improve security.

Supply chain management: Predictive analytics can be used to forecast demand for products and optimize inventory levels. This can help organizations reduce costs and improve efficiency.

Human resources: Predictive analytics can identify employee turnover risk and predict which employees are most likely to leave. This can help organizations develop retention strategies and improve employee satisfaction.

These are just a few examples of the many applications of predictive statistics. In this case, it contains predictive analytics to predict future sales for the next three months for the top five countries with the most sales. Predicting future business growth helps organizations predict the trends of the next three months. The toy company makes more informed decisions and improves business outcomes by leveraging historical data and identifying patterns and trends. Before performing predictive statistics in SAS Studio, first prepare data. This involves several steps, including:

Import toy_company_sales.csv dataset containing 12 months of toy products sold data from 30 countries. Additionally, the dataset has 57 variables and 252076 observations. The toy dataset variables will help find the answer to the hypothesis testing.

Figure 1: the toy_company_sales.csv dataset information.

The CONTENTS Procedure			
Data Set Name	WORK.IMPORT	Observations	252076
Member Type	DATA	Variables	57
Engine	V9	Indexes	0
Created	10/03/2023 10:12:09	Observation Length	464
Last Modified	10/03/2023 10:12:09	Deleted Observations	0
Protection		Compressed	NO
Data Set Type		Sorted	NO
Label			
Data Representation	SOLARIS_X86_64, LINUX_X86_64, ALPHA_TRU64, LINUX_IA64		
Encoding	utf-8 Unicode (UTF-8)		

Clean and prepare the toy_company_sales.csv dataset with code in SAS studio. Clean and prepare your data using data cleaning and standardization techniques. This involves identifying and correcting data errors, inconsistencies, and inaccuracies to improve its quality. To check for missing data in SAS Studio, use the PROC MEANS procedure to count the missing values for each variable in your dataset.

Figure 2: the code checking the missing values in SAS Studio.

```
1 PROC MEANS DATA=WORK.IMPORT NMISS;
2 RUN;
```

Figure 3: Result of the above figure 2 code checking the missing values.

The MEANS Procedure

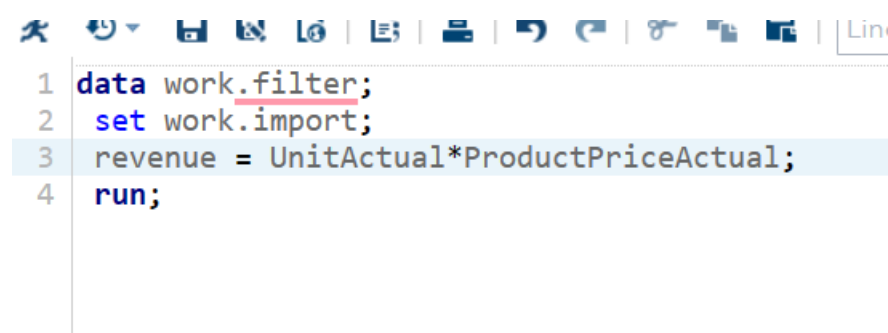
Variable	N Miss
TransactionDate	0
TransactionYear	0
TransactionMonth	0
SalesRepID	0
CustomerLat	0
CustomerLon	0
CustomerDistance	0
FacilityLat	0
FacilityLon	0
FacilityOpeningDate	0
FacilityAge	0
FacilityEmployees	0
FacilityContinentLat	0
FacilityContinentLon	0
FacilityCountryLat	0
FacilityCountryLon	0
FacilityRegionLat	0
FacilityRegionLon	0
FacilityCityLat	0
FacilityCityLon	0
UnitAge	0
UnitStatusCode	0
UnitCapacity	0
UnitTarget	0
UnitActual	0
UnitDiscards	0
ProductPriceTarget	0
ProductPriceActual	0
ProductMaterialCost	0
ProductCostOfSale	0
SalesRepCustomers	0

Messages: 2 User: u61893675

1:20 PM 10/3/2023

The dataset does not contain any missing values. The company wants to forecast future sales, so creating the one-column name 'revenue' will help to use sales trend analysis, which involves analyzing historical revenue data to identify patterns and trends. This can help you detect short-term revenue growth and performance changes and estimate future performance.

Figure 4: add a one-column name 'revenue' in toy company sales data in SAS Studio.



```

1 data work.filter;
2   set work.import;
3   revenue = UnitActual*ProductPriceActual;
4   run;

```

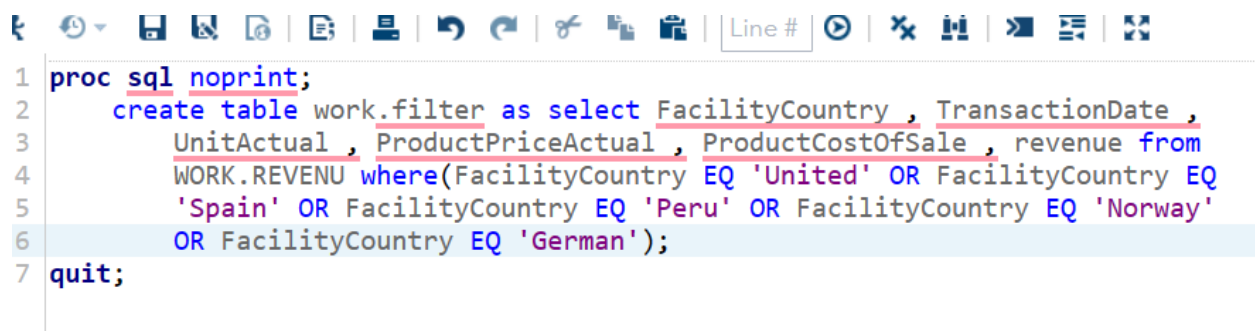
Filter data task helps to create a new table name 'filter' containing five variables

'FacilityCountry', 'TransactionDate', 'UnitActual', 'ProductPriceActual', 'ProductCostOfSale' and

'revenue' with where the condition will take the top five highest revenue countries to order

highest to low United, Spain, Peru, Norway, and German.

Figure 5: code for preparation for this case in SAS Studio.



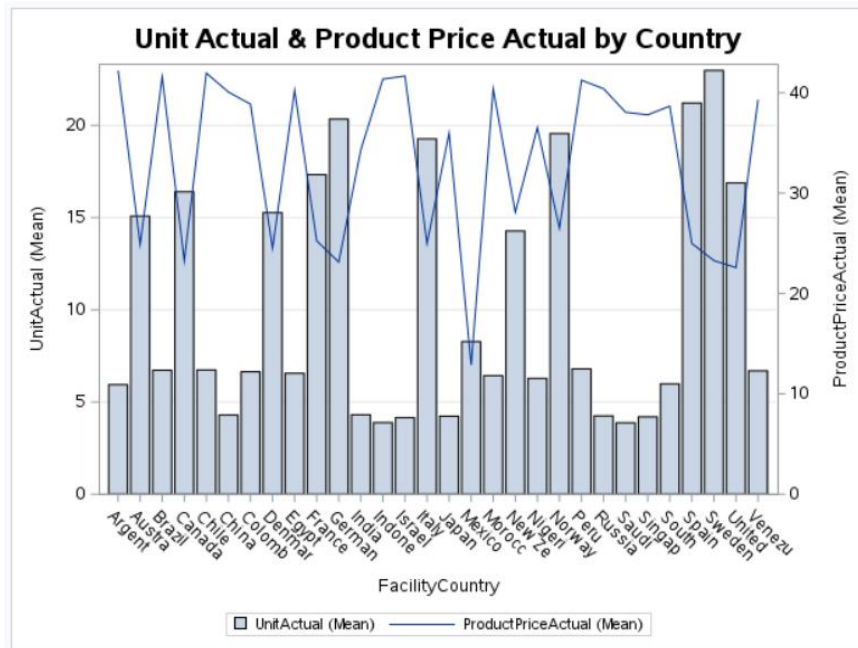
```

1 proc sql noprint;
2   create table work.filter as select FacilityCountry , TransactionDate ,
3     UnitActual , ProductPriceActual , ProductCostOfSale , revenue from
4     WORK.REVENUE where(FacilityCountry EQ 'United' OR FacilityCountry EQ
5     'Spain' OR FacilityCountry EQ 'Peru' OR FacilityCountry EQ 'Norway'
6     OR FacilityCountry EQ 'German');
7 quit;

```

Explore the toy_company_sales.csv dataset using data visualization techniques to identify patterns and trends. This can include creating histograms, scatter plots, and other visualizations to understand the data distribution.

Figure 6: Bar-line chart with UnitActual and ProductPriceActaul variables in SAS Studio.



The five variables 'TransactionDate', 'FacilityCountry', 'UnitActual', 'ProductPriceActual', 'ProductCostOfSale', and 'revenue' with five top companies United, Spain, Peru, Norway, and German will be in predictive statistics processing to find answers for the company's needs.

Figure 7: Worksheet view of the toy company's top five sale countries variables will be used to analyze processing.

Total rows: 199809 Total columns: 6

Rows 301-400

	FacilityCountry	TransactionDate	UnitActual	ProductPriceActual	ProductCostOfSale	revenue
301	United	09/24/2017	1	35	23	35
302	United	09/21/2017	1	7	6	7
303	United	09/18/2017	1	0	0	0
304	United	09/25/2017	1	0	0	0
305	United	09/12/2017	1	22	20	22
306	United	10/04/2017	1	0	0	0
307	United	08/28/2017	1	0	0	0
308	United	11/19/2017	1	0	0	0
309	United	08/15/2017	1	0	0	0
310	United	10/22/2017	1	34	24	34
311	United	11/06/2017	1	0	0	0
312	United	09/03/2017	1	35	33	35
313	United	10/17/2017	1	14	11	14
314	United	09/24/2017	1	39	37	39
315	United	09/25/2017	1	0	0	0

Exploring the SAS Studio task and selecting the Data task, then Characterize Data helps to see descriptive statistics necessary value such as frequency for country or price and cost mean, median, and mode. Option tab selection for this task has both categorical and numerical variables select Frequency table and chart, and Descriptive statistics and Histogram with chosen five variables in figure.

Figure 8: The result of the characterize data window in SAS Studio.

Frequencies for Categorical Variables

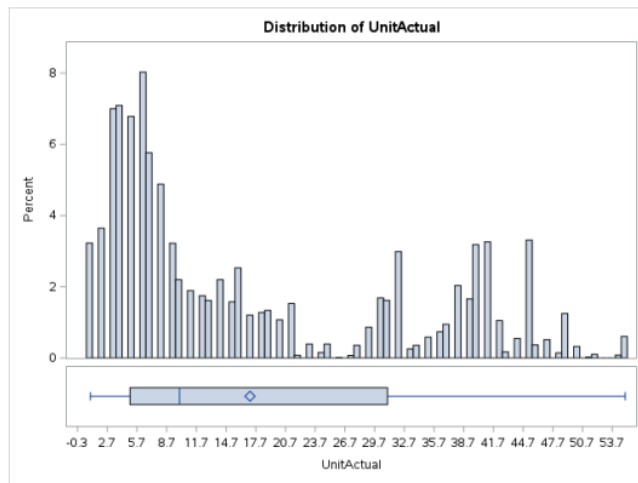
FacilityCountry	Frequency	Percent	Cumulative Frequency	Cumulative Percent
German	5552	2.78	5552	2.78
Norway	3205	1.60	8757	4.38
Peru	3906	1.95	12663	6.34
Spain	13486	6.75	26149	13.09
United	173660	86.91	199809	100.00

Figure 9: Descriptive statistics table and histogram with box plot for the numerical dataset in SAS Studio.

10/4/23, 12:57 PM

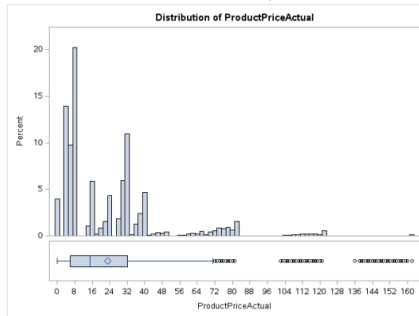
Results: Summary Statistics

Variable	Mean	Std Dev	Minimum	Maximum	N	Skewness	Kurtosis
UnitActual	17.1162310	15.1284331	1.0000000	55.0000000	199809	0.8156666	-0.8161731
ProductPriceActual	23.1853570	24.8596975	0	162.0000000	199809	2.0959029	5.4034949
ProductCostOfSale	19.7139869	19.7837833	0	129.0000000	199809	2.0234375	5.4824596
revenue	211.0775090	165.0289595	0	1539.00	199809	2.6253026	12.2419601



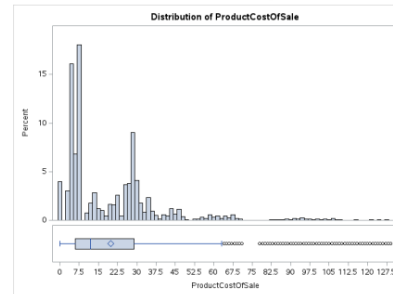
10/4/23, 12:57 PM

Results: Summary Statistics



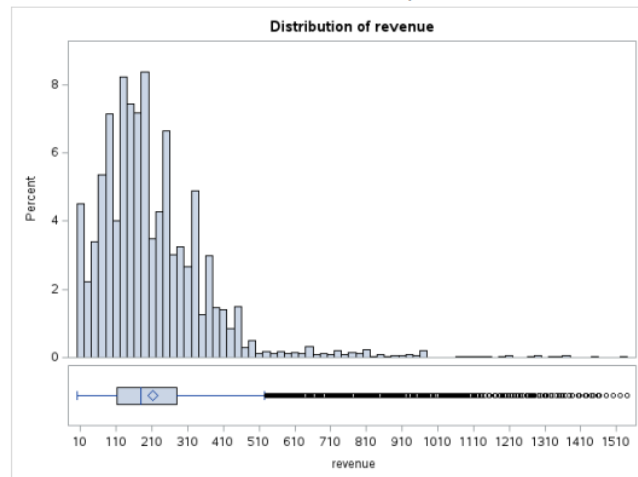
10/4/23, 12:53 PM

Results: Summary Statistics



10/4/23, 12:57 PM

Results: Summary Statistics



The Summary Statistics task outputs show that the revenue's standard deviation is highest at 165, UnitActual at 15.13, ProductPriceActual at 24.86, and ProductCostOfSale at 19.83.

Standard deviation tells, on average, how far each value lies from the mean. A low standard deviation indicates that the values are close to the mean, while a high standard deviation indicates that the values are spread over a broader range.

Kurtosis shows the variable's probability or frequency, which also helps to compare which variable has a heavy distribution tail with three kurtosis types. Revenue has a high kurtosis, and the other two variables, ProductPriceActual and ProductCostOfSale, have a leptokurtic positive value. UnitActual value close to zero is mesokurtic.

Skewness tells whether the dataset has an asymmetric distribution or not that measures three different distributions. Zero skew means the distribution is symmetrical—another negative skew when the number is negative and a positive skew when the number is positive. UnitActual is close to zero, which can be accepted for normal distribution if outliers exceed what we expect. Other variables' skewness is positive, so variables have an outlier on the right side.

Handling Outliers

The toy company data set variables are 'ProductPriceActual,' ProductCostOfSale, 'and 'revenue' have outlier variables. The filter setting needs to change, and three variables have the highest number. Hence, the table to Filter task is set lower than the standard deviation with three variables that would help to remove outliers. The outliers might have vital information, so the case outliers are removed. First, calculate each variable's standard deviation:

For ProductPriceActual variable's standard deviation calculation:

the mean (μ) of 17.1162310,

the standard deviation (σ) of 24.8596975

as $\mu + \sigma = 17.1162310 + 3 * 24.8596975 = 91.6953235$.

For ProductCostOfSale variable's standard deviation calculation:

the mean (μ) of 23.1853570,

the standard deviation (σ) of 19.7837933

as $\mu + \sigma = 23.1853570 + 3 * 19.7837933 = 79.1351732$.

For revenue variable's standard deviation calculation:

the mean (μ) of 211.0775090,

the standard deviation (σ) of 165.0289595

as $\mu + \sigma = 211.0775090 + 3 * 165.0289595 = 706.1643875$.


Figure 10: Code for removing outliers from three variables: revenue, ProductCostOfSale, and ProductPriceActual.

```
proc sql noprint;
  create table work.filter0002 as select * from WORK.FILTER
  where (ProductPriceActual LT 91.6953235 AND ProductCostOfSale LT 79.1351732 OR
  revenue LT 706.1643875);
quit;
```

Line 17, Column 42 UTF-8
Messages: 3 User: u61893675
11:47 AM 10/5/2023

Figure 11: Summary statistics after removing outliers in SAS Studio.

Variable	Mean	Std Dev	Minimum	Maximum	N	Skewness	Kurtosis
ProductPriceActual	22.4229133	23.1801990	0	162.0000000	198400	1.9262289	4.4930365
ProductCostOfSale	19.1021724	18.4355659	0	129.0000000	198400	1.8421903	4.7083236
revenue	206.3907157	155.7064031	0	1539.00	198400	2.6609952	14.0757112



Covariance and Correlation

One of the other causes of the model's performance is multicollinearity. Correlation analysis helps us to know how to relate two variables, as if the high correlation between two variables may negatively affect the predictive result.

Figure 12: Result of Correlation Analysis for toy company's filter table.

4 With Variables:	TransactionDate UnitActual ProductPriceActual ProductCostOfSale
1 Variables:	revenue

Pearson Correlation Coefficients, N = 198400	
	revenue
TransactionDate	0.12957
UnitActual	0.21615
ProductPriceActual	0.48780
ProductCostOfSale	0.48081

Hypothesis Testing

These are the business problems that I will explore, which would be very helpful for business.

Here is a hypothesis question for predicting future sales for the next three months for the top five countries with the most sales:

1. Does the historical sales data for the top five countries with the most sales provide sufficient evidence to predict future sales for the next three months?

Null hypothesis: There is no significant relationship between historical sales data and future sales for the top five countries with the most sales.

Alternative hypothesis: There is a significant relationship between historical sales data and future sales for the top five countries with the most sales.

Figure 13: Result of Two-sample test for top five sales countries with revenue in SAS Studio.



The Kolmogorov-Smirnov test is a nonparametric test that can compare a sample with a reference probability distribution or two samples. The test quantifies the distance between the empirical distribution function of the sample and the cumulative distribution function of the reference distribution or between the practical distribution functions of two samples. The test's null hypothesis is that the representative (s) come from the same distribution as the reference distribution.

The $Pr > |D|$ value in the output of the Kolmogorov-Smirnov test is the p-value for the test. The p-value is the probability of observing a test statistic as extreme as the one computed from the sample data, assuming that the null hypothesis is true. Suppose the p-value is less than or equal to your significance level (usually 0.05). In that case, the three statistics test results show a p-value lower than 0.05, so reject the null hypothesis and conclude that there is a significant difference between the sample(s) and reference distribution. Also, it means that there is enough evidence to reject the null hypothesis. Thus, the alternative hypothesis is that there is a significant relationship between historical sales data and future sales for the top five countries with the most sales.

Predictive Models

The most important thing for predictive models is to have accurate and relevant data. In the case of predicting future sales, it's important to have historical sales data that is representative of the current market conditions and consumer behavior, as a hypothesis test shows the historical data has enough evidence to predict future sales for the top of the five countries for

three months. The data clean and prepared operation helps to see which variable is free from errors, outliers, and missing values, as UnitActual will be in a model variable.

Another critical factor is to choose the suitable model for the data. Several predictive models exist, including regression, time series, and machine learning models. The choice of model depends on the type of data the project has and the research question the project wants to answer.

In that case, it has time series data and a problem to forecast future sales for the top five countries for three months. Time series models are used when the dataset has a time series dataset and wants to predict future values based on past values. These models assume that the future values of the variable depend on its past values and other factors such as seasonality, trends, and cycles. Examples of time series models include ARIMA, exponential smoothing, and state space models. SAS Studio has a Forecasting task that helps create a model to forecast three months of sales for five top sales countries. Each selection in Figure 12-14 on the Modeling and Forecasting task.

Figure 14: Data tab in the Modeling and Forecasting task in SAS Studio.

The screenshot displays the 'DATA' tab in SAS Studio. It includes a 'WORK.FILTER' dropdown, a 'Filter: (none)' indicator, and a 'NOTE' section stating: 'This task requires data in a valid time series format. To prepare your data, run the Time Series Data Preparation task before starting this task.' Below this, the 'ROLES' section shows 'UnitActual' as the dependent variable. The 'ADDITIONAL ROLES' section shows 'TransactionDate' as the time ID. The 'Properties' section includes 'Interval' (set to 'Second'), 'Multiplier' (1), 'Shift' (1), and 'Season length' (4). At the bottom, the 'Group analysis by' section shows 'FacilityCountry'.

Figure 15: Model tab in the Modeling and Forecasting task in SAS Studio.

▼ MODEL

*Forecasting model type:
ARIMA

▼ Model Settings

▼ ARIMA

Autoregressive order (p): 3

Differencing order (d): 1

Moving average order (q): 1

▼ Seasonal ARIMA

Autoregressive order (P): 0

Differencing order (D): 0

Moving average order (Q): 0

☒ Include intercept in model

▼ Plots

Select plots to display:
Selected plots

▼ Series Plots

☐ Autocorrelations plot

☒ Panels of correlation plots

☒ Panels of cross-correlation plots

☐ Inverse-autocorrelations plot

☐ Partial-autocorrelations plot

▼ Residual Plots

☐ Residual autocorrelations plot

☒ Panel of the residual correlation diagnostics

☐ Histogram of the residuals

☐ Residual inverse-autocorrelations plot

☒ Panel of the residual normality diagnostics

☐ Residual partial-autocorrelations

☐ Normal quantile plot of the residuals

☐ Scatter plot of the residuals against time

☐ Ljung-Box white-noise test p-values at different lags

▼ Forecast Plots

☒ One-step-ahead and multistep-ahead forecasts

☒ Multistep-ahead forecasts in the forecast region

Figure 16: Options tab in the Modeling and Forecasting task in SAS Studio.

DATA MODEL **OPTIONS** OUTPUT INFORMATION

▼ FORECAST SETTINGS

Number of periods to forecast: 3

Forecast confidence level:
95%

Number of periods to hold back: 0

▼ OUTLIER DETECTION

☒ Perform outlier detection

To interpret the result of a forecasting model metrics for future sales volume. There are various metrics, such as:

The chi-square test is a statistical test that can be used to evaluate the goodness of fit of a model to the data. A small p-value (usually less than 0.05) indicates that the residuals are not independent and the model does not fit the data well. A large p-value (generally greater than 0.05) indicates that the residuals are independent and the model fits the data well.

The standard error is a statistical measure of variation or uncertainty in a forecasted value. It tells how much the forecasted value will likely vary from the actual value due to random or sampling errors. A low standard error indicates that the forecasted value is expected to be close to the actual value. In contrast, a high standard error indicates that the forecasted value is likely far from the actual value.

The R-square change is the difference between the R-square values of the two models or different categorical data, like models that use the same set of predictors but different target variables. It measures how much the additional predictors in the second model improve the fit of the regression model compared to the first model. This will tell you how much the predictors better explain the future sales volume of one country than the future sales volume of another country. A higher R-square change means that the country's future sales volume is more predictable by the predictors. A lower or negative R-square change means that the country's future sales volume is more predictable by the predictors.

The ACF plot can be used to identify the number of lags that are needed for a time series model. In an ACF plot, each bar represents the size and direction of the correlation. Bars that extend across the blue region are statistically significant. Autocorrelation measures the relationship between a variable's current and past values. An autocorrelation of +1 represents a perfect positive correlation, while an autocorrelation of -1 represents a perfect negative correlation.

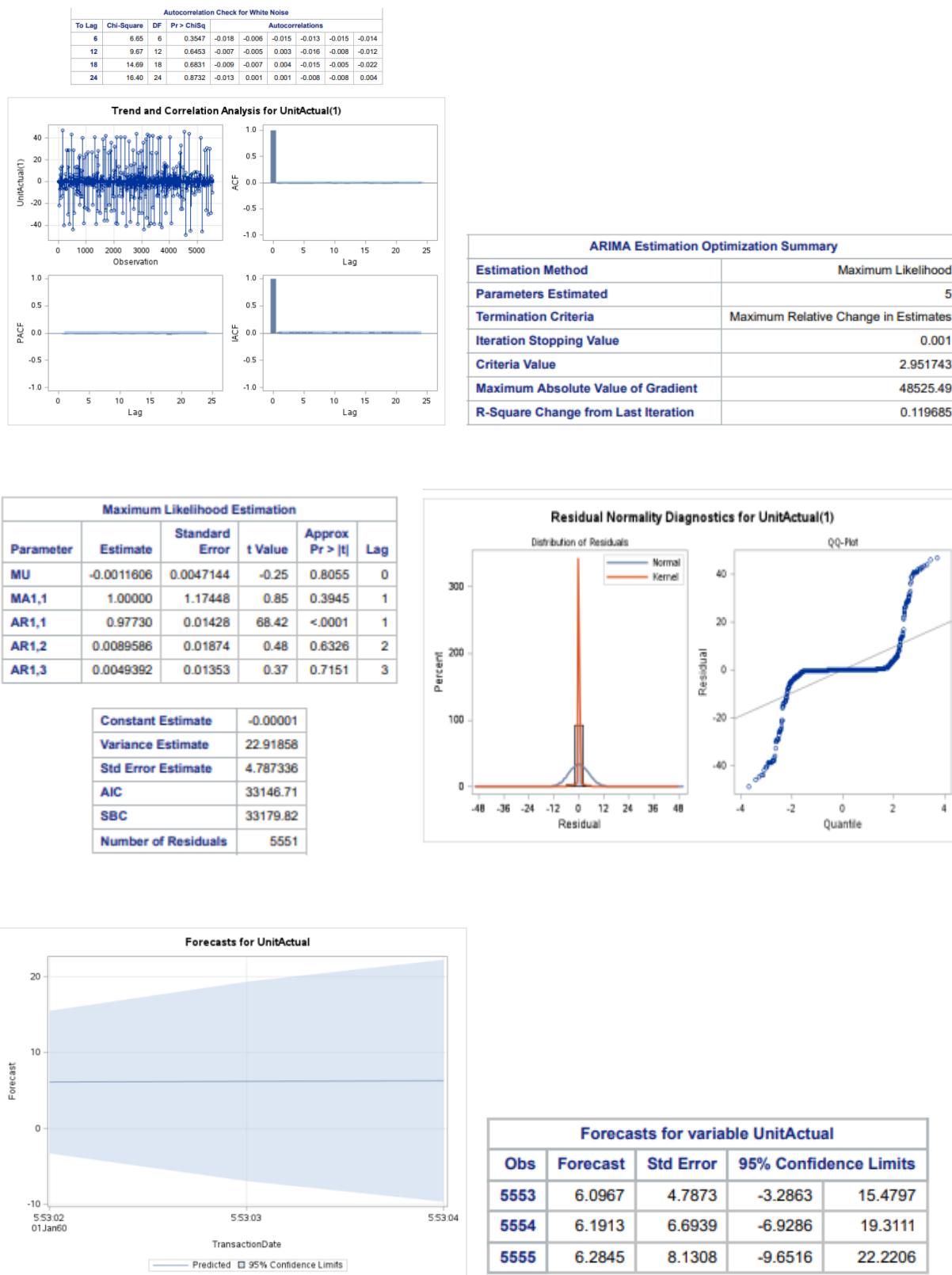
A Q-Q plot is a graphical method used to compare the distribution of a sample with a theoretical distribution, such as the normal distribution. The Q-Q plot compares the quantiles of the sample with the quantiles of the theoretical distribution. If the sample follows the theoretical distribution, the points on the Q-Q plot should fall along a straight line.

The distribution of residuals is a graphical method used to evaluate the goodness of fit of a regression model. Normality: The distribution of residuals should be approximately normal. The model may not fit the data well if the skewed distribution or has heavy tails.

The forecast graph contains trends and patterns so that the forecast graph may show any trends or patterns in the historical and forecasted sales data. These trends and patterns can identify any seasonality, trends, or cycles in the data.

Result of the Forecasting Modeling by each Five Countries

Figure 17: Result of the forecasting model type ARIMA for **German** in SAS Studio.



German's dataset model result of chi-square shows a high p-value that data good fit model. The result of the standard error is low on the Maximum Likelihood Estimation table, the result of a number around 0.01, and the Std Error Estimate at 4.79, so the forecasted value is likely close to the actual value. The table of ARIMA Estimation Optimization Summary has an R-Square value of 0.11, which is low. In predicting future sales, a low R-squared change value indicates that the model may not be a good fit for the data and may not be accurate in predicting future sales that should be compared with another country. The Residual Correlation Diagnostics for UnitActual's result of ACF shows an autocorrelation +1, representing a perfect positive correlation between past and current data. The German Q-Q chart does not have a normal distribution, so the dataset has an outlier. The distribution of residuals has heavy tails that might show a regression model unsuitable for this dataset.

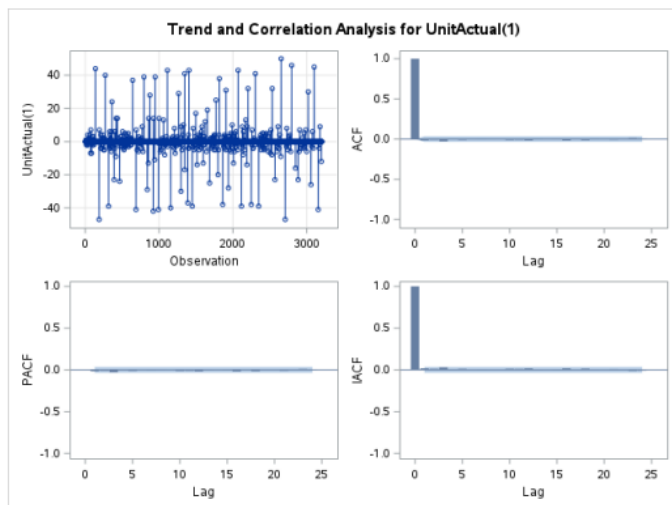
Forecast for UnitActual graph for German shows the x-axis transaction date, and the y-axis shows the unit sales range -10 to 20. The blue shaded area represents the 95% confidence limits, and the black line represents the predicted values for the following primarily three-month per unit sales range in a 95% confidence area, which might be -10 to 20 in German.

Figure 18: Result of the forecasting model type ARIMA for Norway in SAS Studio.

10/5/23, 3:31 PM

Results: Modeling and Forecasting

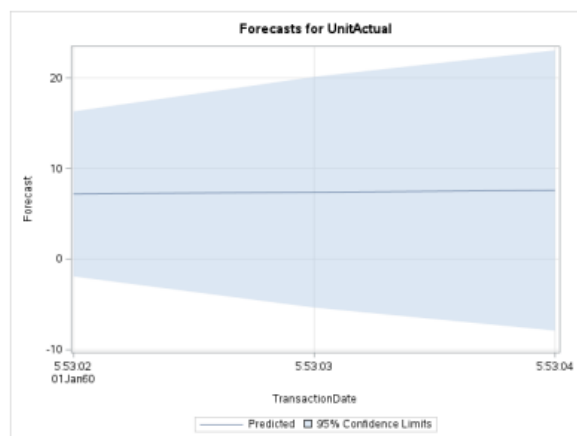
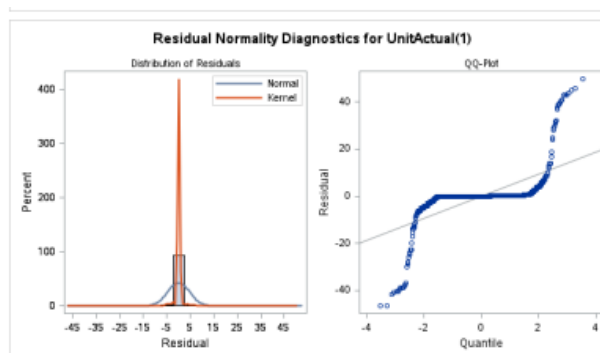
Autocorrelation Check for White Noise									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	4.11	6	0.6623	-0.017	-0.009	-0.028	-0.005	-0.010	-0.000
12	5.86	12	0.9229	-0.005	-0.002	0.000	-0.012	-0.009	-0.017
18	7.71	18	0.9827	0.006	-0.000	-0.004	-0.018	-0.000	-0.014
24	8.66	24	0.9982	0.004	0.002	-0.008	-0.003	0.012	0.008



Maximum Likelihood Estimation					
Parameter	Estimate	Standard Error	t Value	Approx Pr > t	Lag
MU	0.0020058	0.0067007	0.30	0.7647	0
MA1,1	0.99999	0.76092	1.31	0.1888	1
AR1,1	0.97879	0.01838	53.26	<.0001	1
AR1,2	0.0069579	0.02465	0.28	0.7778	2
AR1,3	0.0013286	0.01779	0.07	0.9405	3

Constant Estimate	0.000026
Variance Estimate	21.54273
Std Error Estimate	4.641414
AIC	18937.01
SBC	18967.37
Number of Residuals	3204

ARIMA Estimation Optimization Summary	
Estimation Method	Maximum Likelihood
Parameters Estimated	5
Termination Criteria	Maximum Relative Change in Estimates
Iteration Stopping Value	0.001
Criteria Value	1.696914
Maximum Absolute Value of Gradient	21459.78
R-Square Change from Last Iteration	0.095576



Forecasts for variable UnitActual				
Obs	Forecast	Std Error	95% Confidence Limits	
3206	7.1983	4.6414	-1.8987	16.2953
3207	7.3924	6.4947	-5.3371	20.1219
3208	7.5838	7.8894	-7.8792	23.0468

The Norway dataset ARIMA forecasting result has a few essential metrics: chi-square shows a high $\text{Pr}>\text{ChiSq}$ that data good fit model. The result of the standard error is low on the Maximum Likelihood Estimation table, the result of a number around 0.0, and the Std Error Estimate at 4.64, so the forecasted value is likely close to the actual value. The Residual Correlation Diagnostics for UnitActual's result of ACF shows an autocorrelation +1, representing a perfect positive correlation between past and current data. The Norway Q-Q chart does not have a normal distribution, so the dataset has an outlier. Also, the distribution of residuals has heavy tails that might show a regression model unsuitable for this dataset. The ARIMA Estimation Optimization Summary table has an R-squared value of 0.09, lower than the German dataset.

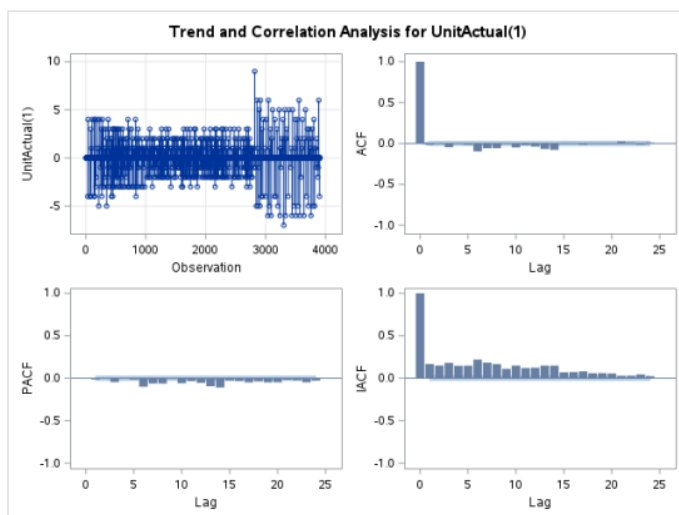
Forecast for UnitActual graph for Norway shows the x-axis transaction date, and the y-axis shows the unit sales range -10 to 20. The blue shaded area represents the 95% confidence limits, and the black line represents the predicted values for the following primarily three-month per unit sales range in a 95% confidence area, which might be -7.88 to 23.05 in Norway.

Figure 19: Result of the forecasting model type ARIMA for **Peru** in SAS Studio.

10/5/23, 3:31 PM

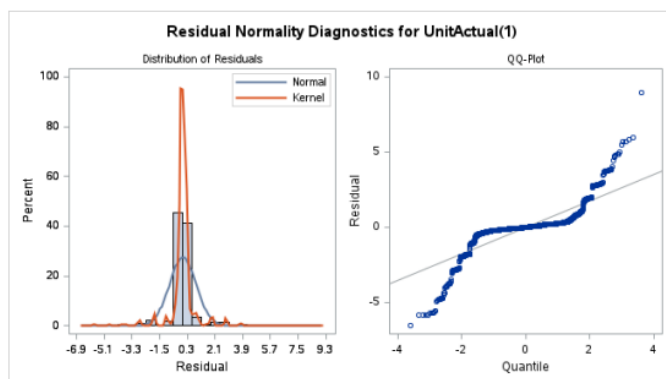
Results: Modeling and Forecasting

Autocorrelation Check for White Noise									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	46.36	6	<.0001	-0.014	-0.006	-0.045	-0.004	-0.014	-0.097
12	88.09	12	<.0001	-0.058	-0.057	0.006	-0.048	-0.019	-0.037
18	130.42	18	<.0001	-0.067	-0.078	0.002	0.003	-0.014	0.000
24	134.79	24	<.0001	-0.006	0.007	0.024	0.012	-0.013	0.011

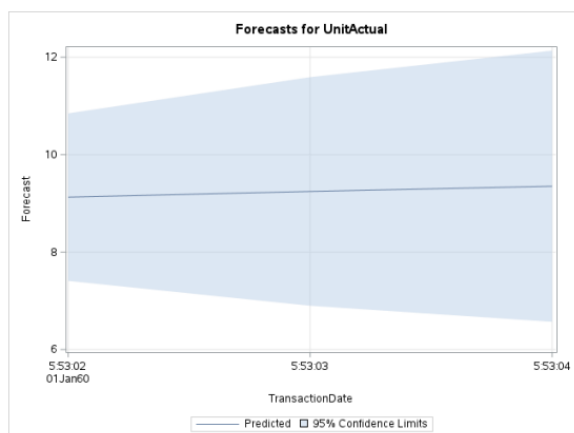


Maximum Likelihood Estimation					
Parameter	Estimate	Standard Error	t Value	Approx Pr > t	Lag
MU	0.0001779	0.0017637	0.10	0.9197	0
MA1,1	0.98144	0.0037601	261.01	<.0001	1
AR1,1	0.91062	0.01636	55.65	<.0001	1
AR1,2	0.0065521	0.02163	0.30	0.7619	2
AR1,3	-0.06657	0.01622	-4.11	<.0001	3

Constant Estimate	0.000027
Variance Estimate	0.768068
Std Error Estimate	0.876395
AIC	10057.55
SBC	10088.9
Number of Residuals	3905



ARIMA Estimation Optimization Summary	
Estimation Method	Maximum Likelihood
Parameters Estimated	5
Termination Criteria	Maximum Relative Change in Estimates
Iteration Stopping Value	0.001
Criteria Value	4.15E-15
Maximum Absolute Value of Gradient	13.26064
R-Square Change from Last Iteration	0.001085
Objective Function	Log Gaussian Likelihood
Objective Function Value	-5023.78
Marquardt's Lambda Coefficient	1E12
Numerical Derivative Perturbation Delta	0.001
Iterations	19
Warning Message	Estimates may not have converged.



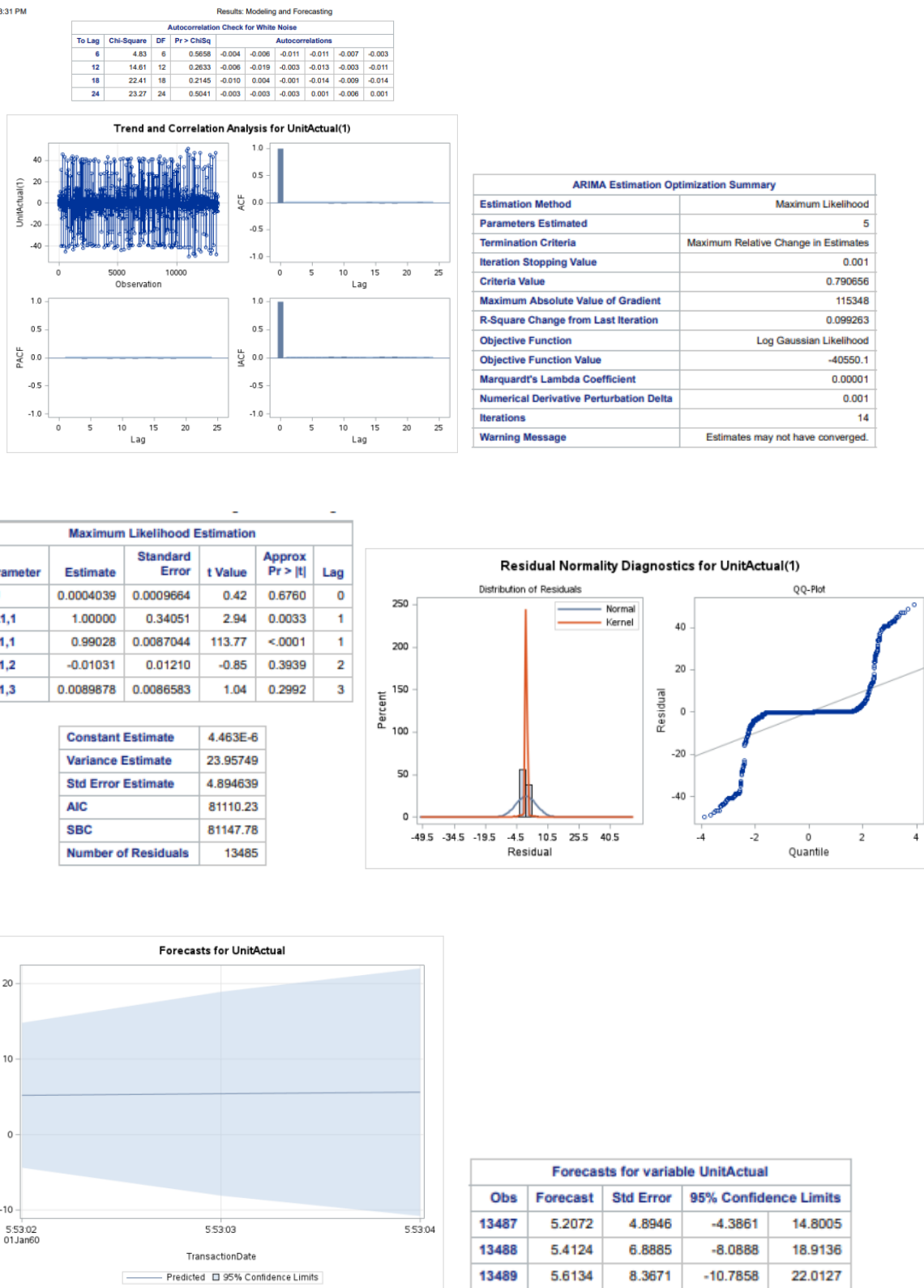
Forecasts for variable UnitActual				
Obs	Forecast	Std Error	95% Confidence Limits	
3907	9.1283	0.8764	7.4106	10.8460
3908	9.2451	1.1963	6.9004	11.5899
3909	9.3524	1.4192	6.5708	12.1340

In the Peru dataset, the chi-square result shows a high value and lower p-value that data good fit model. The result of the standard error is low on the Maximum Likelihood Estimation table, the result of a number around 0.0, and the Std Error Estimate at 0.89, so the forecasted value is too close to the actual value. The Residual Correlation Diagnostics for UnitActual's result of ACF shows an autocorrelation +1, representing a perfect positive correlation between past and current data. The Norway Q-Q chart does not have a normal distribution, so the dataset has an outlier. Also, the distribution of residuals has heavy tails that might show a regression model unsuitable for this dataset. The ARIMA Estimation Optimization Summary table has an R-squared change value of 0.001, lower than the Norway dataset.

Forecast for UnitActual graph for Peru shows the x-axis transaction date, and the y-axis shows the unit sales range 6 to 12. The blue shaded area represents the 95% confidence limits, and the black line represents the predicted values for the following primarily three-month per unit sales range in a 95% confidence area, which might be 6.5 to 12.13 in Peru.

Figure 20: Result of the forecasting model type ARIMA for **Spain** in SAS Studio.

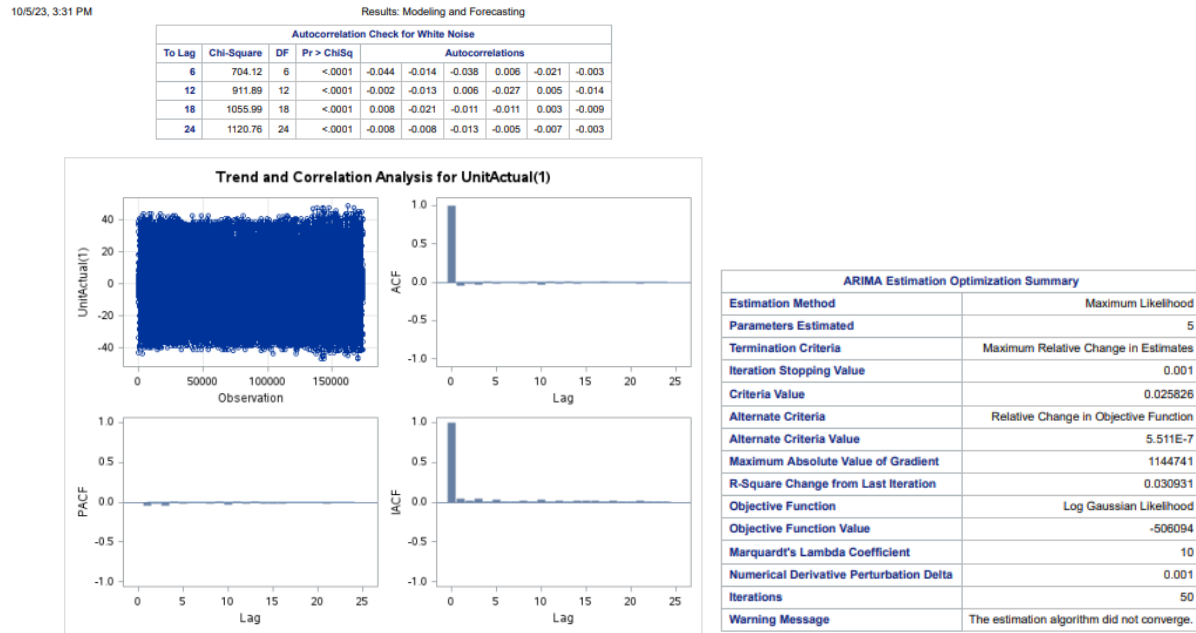
10/5/23, 3:31 PM



The Spain dataset ARIMA forecasting result has a few essential metrics: chi-square shows a high $P > \chi^2$ around 0.21-0.56 that data good fit model. The result of the standard error is low on the Maximum Likelihood Estimation table, the result of a number around 0.0, and the Std Error Estimate at 4.89, so the forecasted value is likely close to the actual value. The Residual Correlation Diagnostics for UnitActual's result of ACF shows an autocorrelation +1, representing a perfect positive correlation between past and current data. The Norway Q-Q chart does not have a normal distribution, so the dataset has an outlier. Also, the distribution of residuals has heavy tails that might show a regression model unsuitable for this dataset. The ARIMA Estimation Optimization Summary table has an R-squared value of 0.09, the same as the Peru dataset.

Forecast for UnitActual graph for Norway shows the x-axis transaction date, and the y-axis shows the unit sales range -10 to 20. The blue shaded area represents the 95% confidence limits, and the black line represents the predicted values for the following primarily three-month per unit sales range in a 95% confidence area, which might be -10.78 to 22.01 in Spain.

Figure 21: Result of the forecasting model type ARIMA for **United** in SAS Studio.



In the United dataset, the chi-square result shows a high value and lower p-value that data good fit model. The Residual Correlation Diagnostics for UnitActual's result of ACF shows an autocorrelation +1, representing a perfect positive correlation between past and current data.

The ARIMA Estimation Optimization Summary table has an R-squared change value of 0.030.

Forecasting for the following three months for United as the Trend and Correlation multiple graphs show the observation plot shows the difference between the actual and predicted sales units for each observation. This is also known as the residual or error of the prediction model.

Recommendation

Pr>ChiSq for five different companies' future sales volume needs to compare the p-values of their ARIMA models. Peru has a p-value of 0.0001, the lowest one, and Norway has a p-value of 0.66 is the highest one then conclude that Peru's model has significant autocorrelation in the

residuals and does not fit the data well. In contrast, country Norway's model has no autocorrelation in the residuals and fits the data well. This means the ARIMA model predicts country Norway's future sales volume more than Peru's.

To compare the future sales volume of five countries, the R-square changes by subtracting the R-square value of the model for one country from the R-square value of the model for another country. This will tell how much the predictors better explain the future sales volume of one country than that of others. A higher R-square change means that Germany at 0.1196 future sales volume is more predictable by the predictors. A lower or negative R-square change means that Peru at 0.001 future sales volume is more predictable by the predictors.

Country Name	chi-square	Pr>ChiSq	standard error estimate	R-square change
German	6.65	0.35	4.78	0.1196
Norway	4.11	0.66	4.64	0.0955
Peru	46.36	0.0001	0.87	0.001
Spain	4.38	0.56	4.89	0.0992
United	704.12	0.001		0.0309

References

- Halton, C. (2023). *Predictive Analytics: Definition, Model Types, and Uses*.
<https://www.investopedia.com/terms/p/predictive-analytics.asp>
- Cody, R.(2021). *A Gently Introduction to Statistics Using SAS Studio in the Cloud*. SAS Institute.
 ISBN: 9781954844476.
- Frimodig, B. (2023). *Chi-Square (X^2) Test & How To Calculate Formula Equation*.
<https://www.simplypsychology.org/chi-square.html>
- Analysis INN.com (2020). *The meaning of R, R Square, Adjusted R Square, R Square Change and F Change in a regression analysis*. <https://www.analysisinn.com/post/the-meaning-of-r-r-square-adjusted-r-square-r-square-change-and-f-change-in-a-regression-analysis/>
- Minitab.com (n.d.). *Interpret the key results for ARIMA*. <https://support.minitab.com/en-us/minitab/21/help-and-how-to/statistical-modeling/time-series/how-to/arma/interpret-the-results/key-results/?SID=117600&SID=117600>
- SAS.com (n.d.). *Time Series Modeling and Forecasting Using SAS Studio*.
<https://video.sas.com/detail/video/4414522302001/time-series-modeling-and-forecasting-using-sas-studio>
- SAS.com (n.d.). *Modeling and Forecasting Task*.
<https://support.sas.com/documentation/cdl/en/webeditorug/68254/HTML/default/viewer.htm#p0grsnazj78fjwn10jtti95r2itr.htm>