Project 4 Retail Analysis

Submitted by Pravin Wagh

Case Study

Problem	Forecast the sales based on the independent variables such as Profit, Quantity, Marketing cost, and Expenses using the regression model.The dataset is maintained for the Retail Analysis, and it has records of both independent and dependent variables.
Analysis	Steps to perform Retail Analysis: • Import the required dataset • Perform descriptive statistics for the dataset • Check the significance of independent variables • Create a new data set with exponential, cube, squared, and log values for each variable • Perform regression test • Print the output dataset
Dataset	Project 04_Retail Analysis_Dataset.xlsx Columns :{Order_ID, Products, Sales, Quantity, Discount, Profit, Shipping_Cost}
Attributes (Independent Variables)	Quantity | Discount | Profit | Shipping_Cost
Category	Order_ID | Products
Target (Dependent Variable)	Sales

Summary

The objective of analysis is to forecast Sales based on the independent variable. The nature of data is a continuous data and the distribution is Normal Distribution.

Feature of the Dataset:

The Dataset contains 30 records.

There are 9 Distinct Products

The Datatype is float.

Discounts offered ranges from 0.01 to 0.05

Step 1

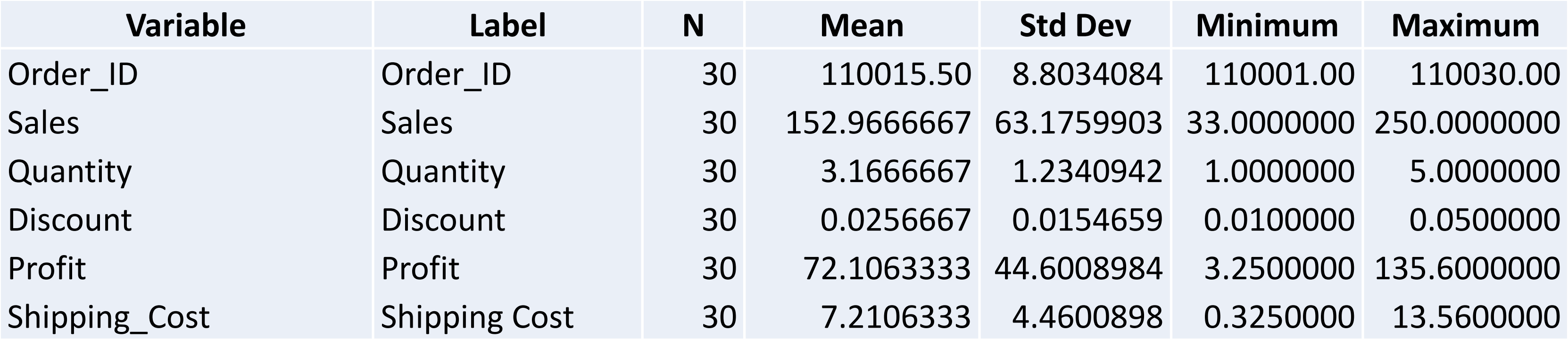
Importing Dataset in SAS

```
FILENAME REFFILE '/folders/myfolders/Project 04_Retail
Analysis_Dataset.xlsx';
PROC IMPORT DATAFILE=REFFILE
            DBMS=XLSX
            OUT=WORK.Retail;
            GETNAMES=YES;
RUN;
```

Step 2: Descriptive Statistics

Finding Mean, Standard Deviation, Minimum and Maximum Values of Variables

```
Proc Means Data=Retail;
            title 'Mean Values of Variables';
Run;
```



Variable	Label	N	Mean	Std Dev	Minimum	Maximum
Order_ID	Order_ID	30	110015.50	8.8034084	110001.00	110030.00
Sales	Sales	30	152.9666667	63.1759903	33.0000000	250.0000000
Quantity	Quantity	30	3.1666667	1.2340942	1.0000000	5.0000000
Discount	Discount	30	0.0256667	0.0154659	0.0100000	0.0500000
Profit	Profit	30	72.1063333	44.6008984	3.2500000	135.6000000
Shipping_Cost	Shipping Cost	30	7.2106333	4.4600898	0.3250000	13.5600000

Finding Frequency of Product Category

```
Proc Freq DATA=Retail;
        Table Products;
        title 'Products Frequency';
Run;
```

Products	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Product1	4	13.33	4	13.33
Product2	4	13.33	8	26.67
Product3	4	13.33	12	40.00
Product4	4	13.33	16	53.33
Product5	3	10.00	19	63.33
Product6	3	10.00	22	73.33
Product7	3	10.00	25	83.33
Product8	3	10.00	28	93.33
Product9	2	6.67	30	100.00

Detail Data Analysis by applying Univariate Function

By applying Univariate function we understand basic statistical features like Mean, Median, Mode, Std deviation, Variance, Range, Quantiles distribution and Extreme Observation in Sales Dataset,

By plotting Histogram we try to understand the spread of the data.

```
Proc Univariate data=Retail;
        var Sales;
        Histogram;
        title 'Descriptive Statistics';
Run;
```

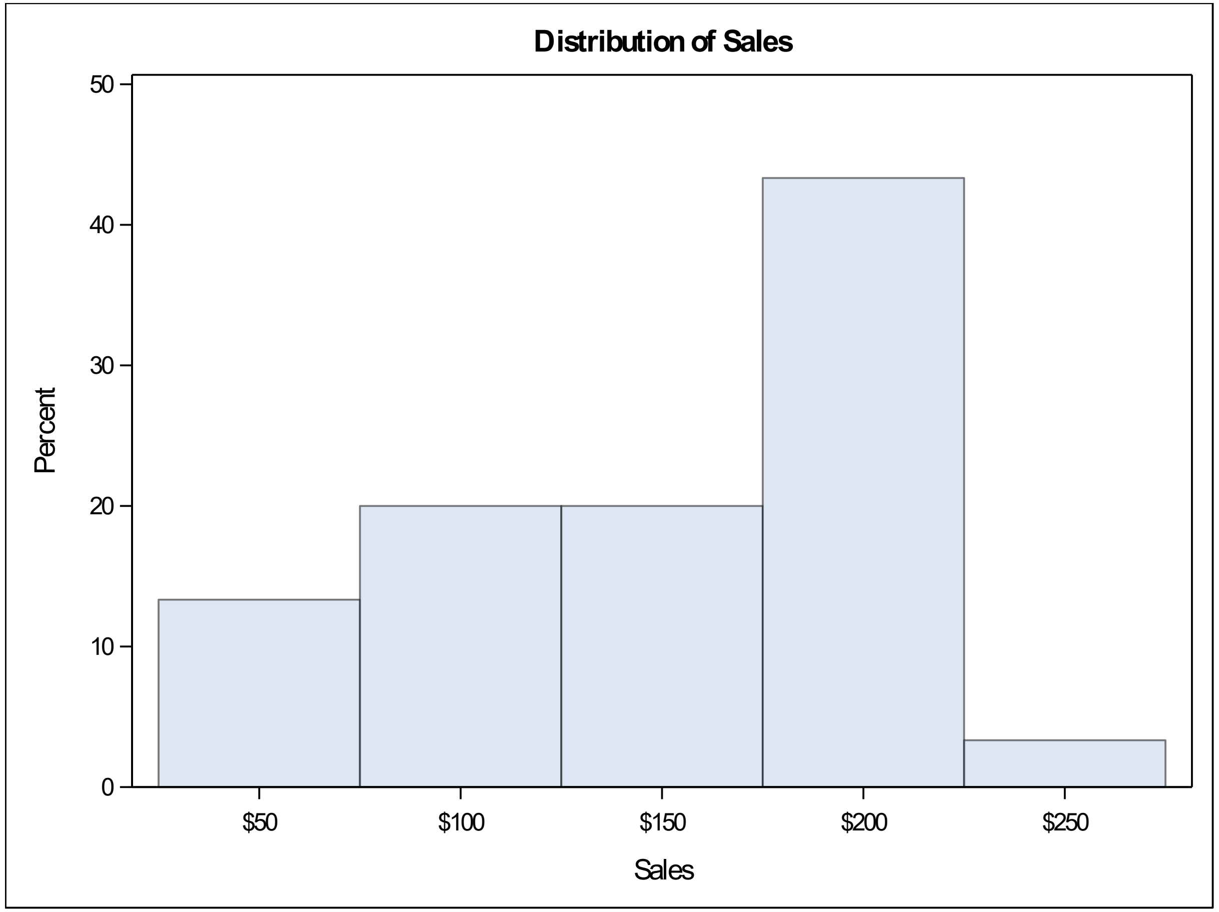
Detail Data Analysis by applying Univariate Function (….contd)

Moments			
N	30	Sum Weights	30
Mean	152.966667	Sum Observations	4589
Std Deviation	63.1759903	Variance	3991.20575
Skewness	-0.3528421	Kurtosis	-1.063775
Uncorrected SS	817709	Corrected SS	115744.967
Coeff Variation	41.3004948	Std Error Mean	11.534305

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	13.26189	Pr > |t|	<.0001
Sign	M	15	Pr >= |M|	<.0001
Signed Rank	S	232.5	Pr >= |S|	<.0001

Quantiles (Definition 5)	
Level	Quantile
100% Max	250
99%	250
95%	222
90%	221
75% Q3	220
50% Median	149
25% Q1	104
10%	65
5%	33
1%	33
0% Min	33

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
33	19	220	27
33	7	220	29
65	23	222	4
65	11	222	16
83	21	250	8



Check the significance of independent variables

Through Anova test we try to understand significance of independent and a dependent variable. We try to understand the dependency of Sales vs Discount. That is, does Discount plays role in increase of sales?

Hypothesis

Ho :=> Sales is relational with the discount

AND

H1 :=> Sales is has no relation with the discount

```
Proc Anova Data=Retail;
                    title 'Annova';
                    Class Discount;
                    Model Sales = Discount;
Run;

Proc PLot;
          Plot Sales * Discount;
Run;
```

Check the significance of independent variables

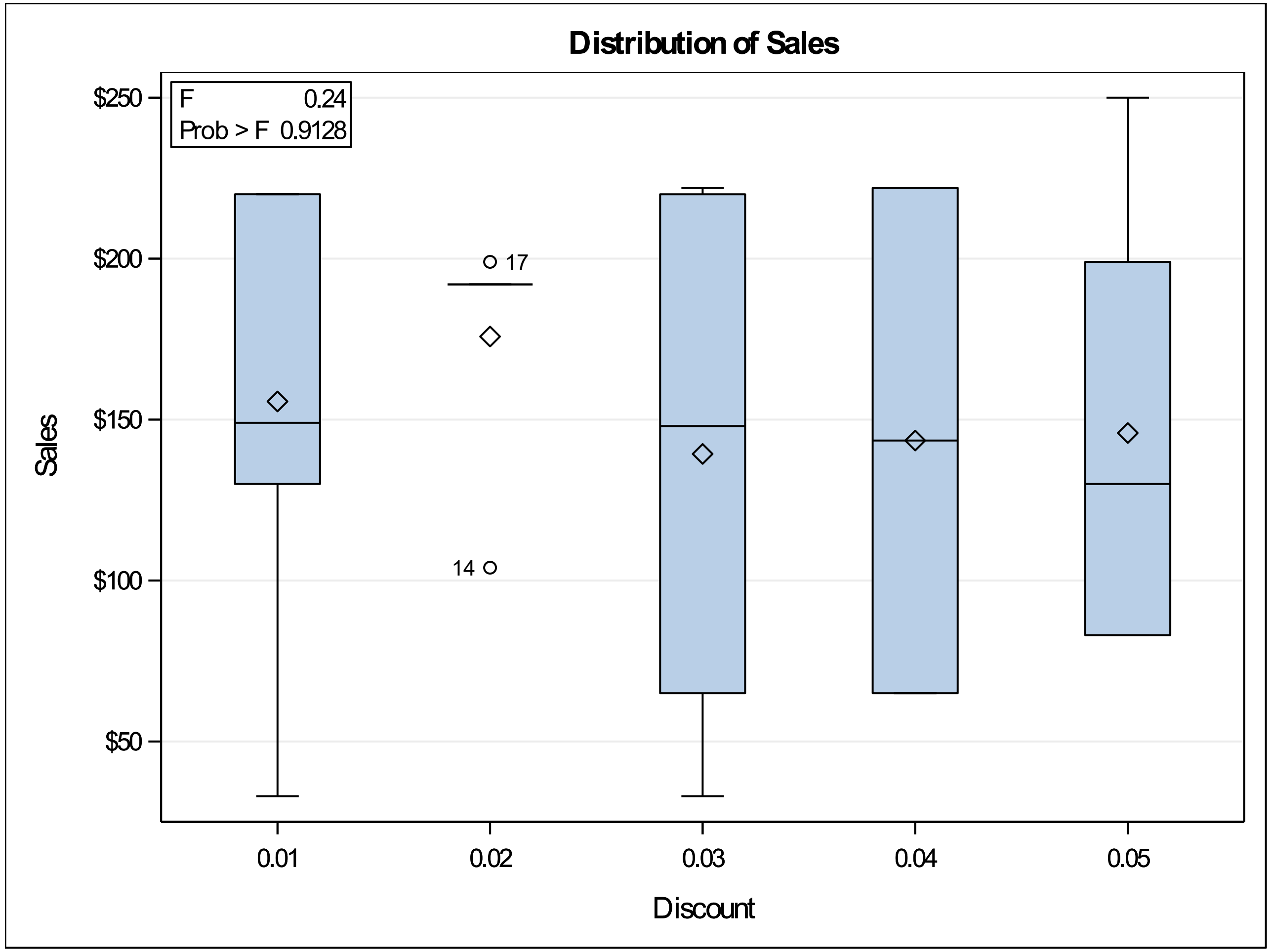
Class Level Information		
Class	Levels	Values
Discount	5	0.01 0.02 0.03 0.04 0.05

Number of Observations Read	30
Number of Observations Used	30

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	4284.9545	1071.2386	0.24	0.9128
Error	25	111460.0121	4458.4005		
Corrected Total	29	115744.9667			

R-Square	Coeff Var	Root MSE	Sales Mean
0.037021	43.65085	66.77125	152.9667

Source	DF	Anova SS	Mean Square	F Value	Pr > F
Discount	4	4284.954545	1071.238636	0.24	0.9128



Create a new data set with exponential, cube, squared, and log values for each variable

```
Data FinalNewRetail;
        Set finalnewretail;
        exp_Discount = EXP(Discount);
        exp_Profit = EXP(Profit);
        exp_ShipCost = EXP(Shipping_Cost);
        lny_Discount = log(Discount);
        lny_Profit = log(Profit);
        lny_ShipCost = log(Shipping_Cost);
        cube_Discount = Discount**3;
        cube_Profit = Profit**3;
        cube_ShipCost = Shipping_Cost**3;
        sq_Discount = Discount**2;
        sq_Profit = Profit**2;
        sq_ShipCost = Shipping_Cost**2;
Run;
```

Result: => *FinalProject_Expo_Log_Cube.xlsx*

Perform regression test

As from the Data one can understand the data is continuous and relational through ANOVA. With Regression test we can analyze the fitness of the data with respect to linear regression line.
In order to conduct regression test we are considering 2 independent variable Quantity and Discount and dependent variable Sales.
First Discount in percentage is converted in real number (amount).
We build Hypothesis
Ho : Relation between Sales and Quantity and Discount is strongly correlated and fit the regression line.
H1 : Relation between Sales and Quantity and Discount is not strongly correlated and doesn’t fit the regression line.

As Pr>F value is less than 0.05 hence we accept Null Hypothesis and model is linearly proportion.

```
Proc Sql;
        Create table RetailReg as
        Select Order_ID, Products, Sales, Quantity, Discount, Profit, Shipping_Cost from Retail;
Quit;

Data RegDisCalc;
        Set RetailReg;
        CalcDisc = Sales * Discount;
Run;

Proc Reg Data=RegDisCalc;
        Model Sales = Quantity CalcDisc;
Run;
```

Perform regression test

Number of Observations Read	30
Number of Observations Used	30

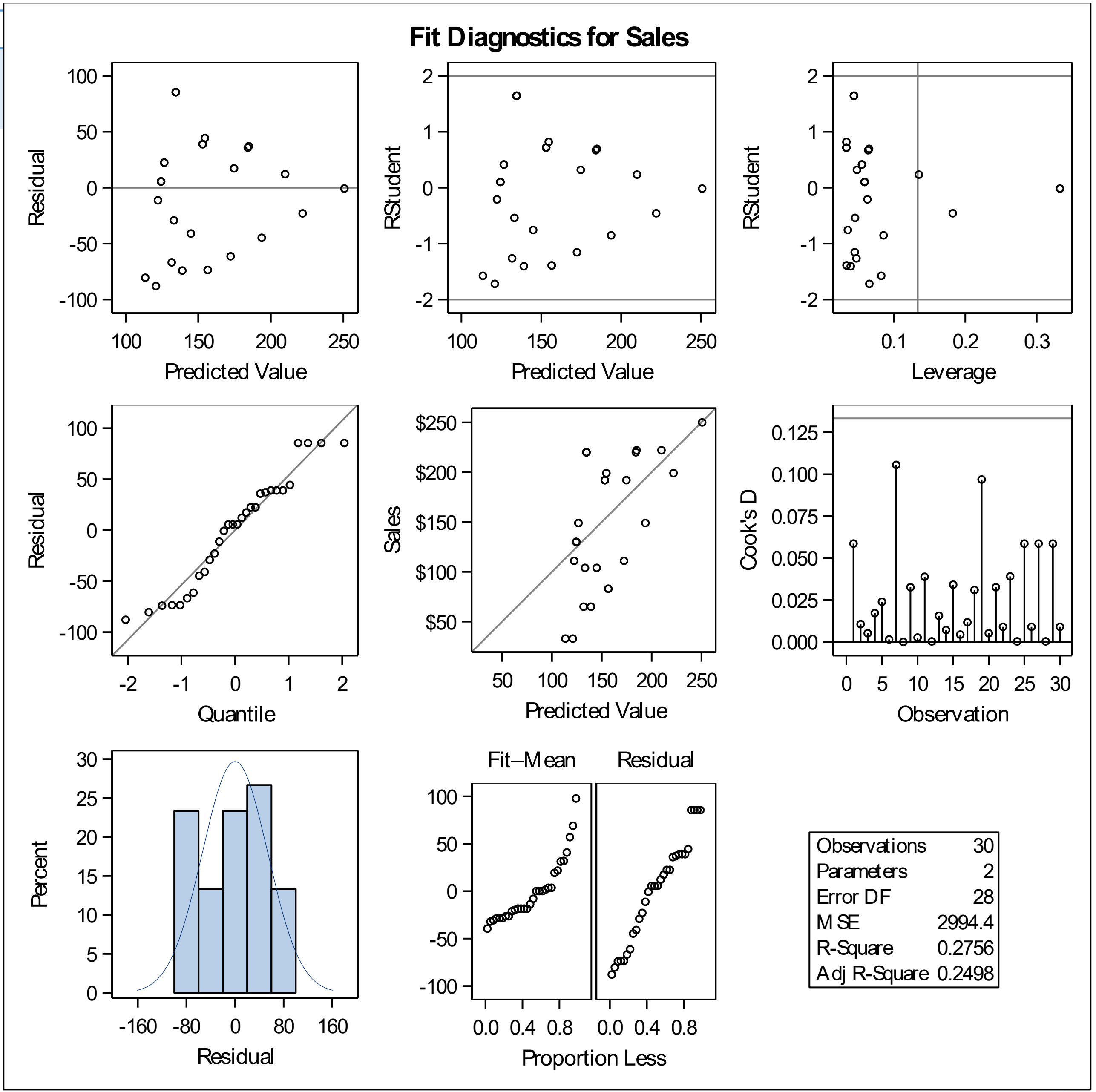
Analysis of Variance

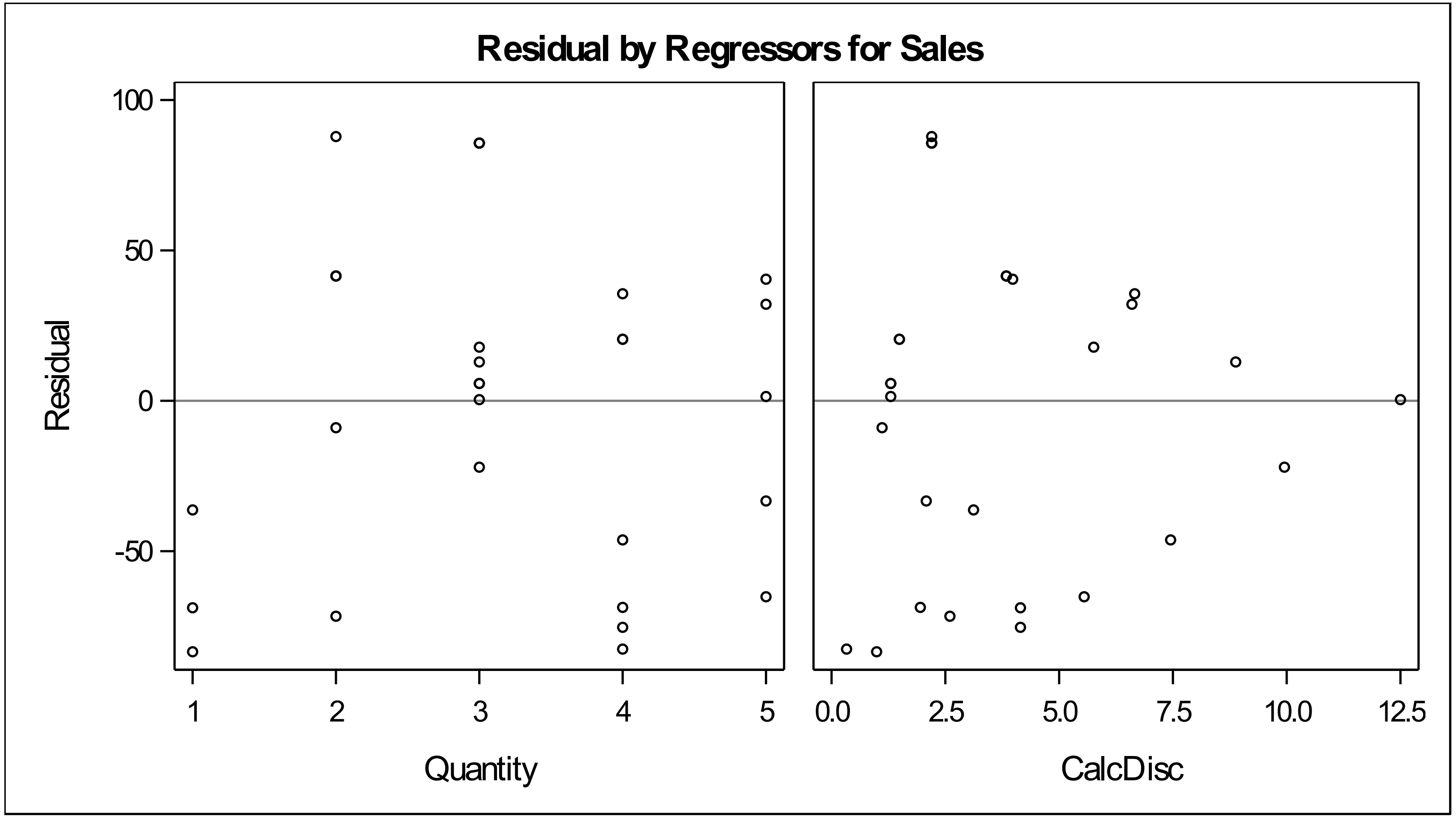
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	32108	16054	5.18	0.0124
Error	27	83637	3097.67457		
Corrected Total	29	115745			

Root MSE	55.65676	R-Square	0.2774
Dependent Mean	152.96667	Adj R-Sq	0.2239
Coeff Var	36.38489		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	103.19096	30.50710	3.38	0.0022
Quantity	Quantity	1	2.16680	8.40847	0.26	0.7986
CalcDisc		1	11.19403	3.52766	3.17	0.0037




Residual by Regressors for Sales
Residual
100
50
0
-50
1
2
3
4
5
Quantity
0.0
2.5
5.0
7.5
10.0
12.5
CalcDisc