Erik Thompson
DSCI 449
Spring 2024 -HW 1
Data Examination

Research Problem

This study will focus on the ratings of technical support offered by HBAT industries as reported by respondents or purchasing managers based on the length of relationship. A random and independent sample size of 100 with a confidence level of 95% will be used in the examination. Three values will be used to categorize the customer types, 1= less than 1 year; 2=between 1 and 2 years; 3=longer than 5 years.

My variable, Tech_support is approximately normally distributed. These findings are supported by mean of 5.37 < 5.4 median, indicating nearly symmetrical distribution. The skew is slightly negative, but is nearly symmetrical with a skewness value of -0.20, and a kurtosis value of -0.55 indicating a distribution that is slightly flat, but still normal. Shapiro-Wilk indicates normality with W=0.9863 and p-value of 0.39 > α of .05. There is a positive linear relationship in the normal probability plot, the whiskers of the box plot are similar in length, with the mean and median close to the center, and the histogram appears to be multi-modal validating a normal distribution. There are no outliers in the box plot, no data points above or below the upper and lower fence. My variable does not require transformation because normality can be assumed without any adjustments to the model.

Regression output shows no outliers as defined as standardized residuals with absolute values > 3. The largest is observation 31 with a student residual of 2.16. The absence of bivariate outliers is supported by the lack of values outside the ellipse in the scatterplot. Univariate analysis of residuals is shown to be normal with a mean close to zero, nearly symmetric skewness with a value of -0.18, and a kurtosis value of -0.62.Shapiro-Wilk indicates normality with W=0.9876 and a p-value of 0.48 > α of .05. Regression output shows no relationship between the variables with F(1,98)=0.92 and a P-value of 0.3387. Homoscedasticity is valid because there is no apparent pattern in the scatterplot or residual plots. Lots of scatter, but linear relationship is possible. Homogeneity of variances appears to be valid, Levene's test is not significant, F(2,97)=0.73 and P-value 0.3841. Boxplots appear to be similar among the 3 customer type values. There are no visible outliers in the boxplots for the 3 customer types.

Research Study Background Information

The HBAT database has one hundred observations with 23 variables and is assumed to be based on surveys of HBAT customers completed on a secure Web site managed by an established business research company. The research company contacts purchasing managers and encourages them to participate. Managers log onto the Web site and complete the survey. The survey results are supplemented by other information, compiled, and stored in HBAT's data warehouse and accessible through its decision support system.

The SAS System

The MEANS Procedure

HBAT sells paper products to two market segments: the newsprint industry and the magazine industry. Paper products are sold to these market segments either directly to the customer or indirectly through a broker. The survey questionnaire consisted of two sections. The first section addressed perceptions of HBAT's performance on thirteen attributes. These attributes were developed from focus groups, pretested, and have been used in previous studies. They are considered the most influential in the selection of suppliers in the paper industry. Respondents were purchasing managers of firms buying from HBAT, and they rated HBAT on each of the 13 attributes using a 0-10 scale, where 10 = "excellent" and 0 = "poor". The second section relates to purchase outcomes and business relationships (e.g., satisfaction with HBAT and whether the firm would consider a strategic alliance or partnership with HBAT). A third type of information was available from HBAT's data warehouse and includes information such as size of customer and length of purchase relationship.

Of the twenty-three variables 17 were metric (quantitative). Thirteen variables measured perceptions of HBAT, four variables measured purchase outcomes (e.g., purchase relationships with HBAT).

Variable	N	N Miss	Mean	Std Dev	Minimum	Maximum
Prod_Qual	100	0	7.8100000	1.3962793	5.0000000	10.0000000
Ecommerce	100	0	3.6720000	0.7005164	2.2000000	5.7000000
Tech_support	100	0	5.3650000	1.5304568	1.3000000	8.5000000
Complaint	100	0	5.4420000	1.2084032	2.6000000	7.8000000
Adv	100	0	4.0100000	1.1269428	1.9000000	6.5000000
Prod_Line	100	0	5.8050000	1.3152850	2.3000000	8.4000000
Sales_Image	100	0	5.1230000	1.0723198	2.9000000	8.2000000
Pricing	100	0	6.9740000	1.5450553	3.7000000	9.9000000
Warranty	100	0	6.0430000	0.8197382	4.1000000	8.1000000
New_Prod	100	0	5.1500000	1.4930479	1.7000000	9.5000000
Ordering	100	0	4.2780000	0.9288398	2.0000000	6.7000000
Price_Flex	100	0	4.6100000	1.2060035	2.6000000	7.3000000
Del_Speed	100	0	3.8860000	0.7344372	1.6000000	5.5000000
Satis	100	0	6.9180000	1.1918393	4.7000000	9.9000000
Recommend	100	0	7.0200000	1.0433048	4.6000000	9.9000000
Future_Purch	100	0	7.7130000	0.9361057	5.5000000	9.9000000
Usage_level	100	0	58.4000000	8.8608776	37.1000000	77.1000000

Tech_support has a mean of 5.37, falling in the middle of the 0-10 scale. This suggests and average user experience that is neither poor or excellent. Tech_support does have the lowest minimum value of 1.3, showing that out of the respondents that rated HBAT, Tech_support received the worst score. There were several variables that received a rating above 9, so the maximum rating of 8.5 for Tech_support is not an exceptional score.

The SAS System

The FREQ Procedure

Customer_type	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	32	32.00	32	32.00
2	35	35.00	67	67.00
3	33	33.00	100	100.00

Industry_type	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	52	52.00	52	52.00
1	48	48.00	100	100.00

Firm_size	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	49	49.00	49	49.00
1	51	51.00	100	100.00

Region	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	39	39.00	39	39.00
1	61	61.00	100	100.00

Distribution	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	57	57.00	57	57.00
1	43	43.00	100	100.00

Partner	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	55	55.00	55	55.00
1	45	45.00	100	100.00

The SAS System

The UNIVARIATE Procedure
Variable:
Tech_support

Moments			
N	100	**Sum Weights**	100
Mean	5.365	**Sum Observations**	536.5
Std Deviation	1.53045679	**Variance**	2.34229798
Skewness	-0.2032586	**Kurtosis**	-0.5482262
Uncorrected SS	3110.21	**Corrected SS**	231.8875
Coeff Variation	28.5266876	**Std Error Mean**	0.15304568

Examination of 'Tech_support'

The metric variable 'Tech_support' shows a slightly negative skew, but is nearly symmetrical with a skewness value of -0.20.

A kurtosis value of -0.55 indicates that the distribution is slightly flat, but still a normal distribution.

The mean of 5.37 < 5.4 median also indicating a nearly symmetrical distribution.

The data is multi-modal with a mode of 4.6 for the smallest of 5 modes with a count of 4.

Basic Statistical Measures			
Location		**Variability**	
Mean	5.365000	**Std Deviation**	1.53046
Median	5.400000	**Variance**	2.34230
Mode	4.600000	**Range**	7.20000
		Interquartile Range	2.45000

Note: The mode displayed is the smallest of 5 modes with a count of 4.

Location Test

H_0: μ=0
H_1: μ≠0

t(99) = 35.05
A [p-value below .0001] < α of .05. Reject the null. The mean of 'Tech_support' is statistically different from 0. It is much greater than 0.

Tests for Location: Mu0=0				
Test	**Statistic**		**p Value**	
Student's t	**t**	35.05489	**Pr > |t|**	<.0001
Sign	**M**	50	**Pr >= |M|**	<.0001
Signed Rank	**S**	2525	**Pr >= |S|**	<.0001

Normality Test

H_0: Data is normal.
H_1: Data is not normal.

W=0.9863
A p-value of 0.39 > α of .05. Fail to reject the null. The variable, 'Tech_support' is normally distributed.

Tests for Normality				
Test	**Statistic**		**p Value**	
Shapiro-Wilk	**W**	0.986264	**Pr < W**	0.3900
Kolmogorov-Smirnov	**D**	0.060152	**Pr > D**	>0.1500
Cramer-von Mises	**W-Sq**	0.051364	**Pr > W-Sq**	>0.2500
Anderson-Darling	**A-Sq**	0.370975	**Pr > A-Sq**	>0.2500

The SAS System

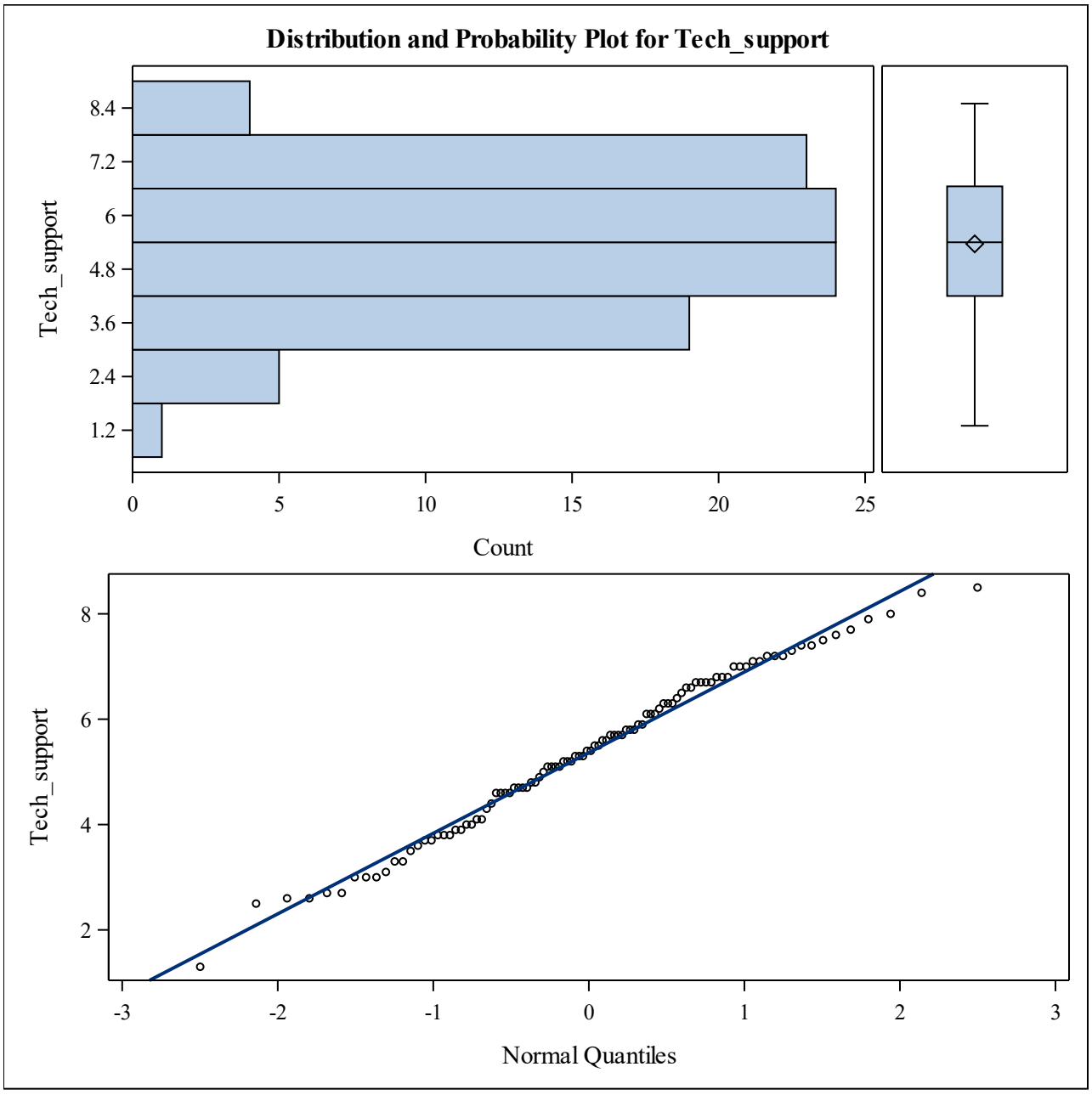
The UNIVARIATE Procedure
Variable:
Tech_support

Quantiles (Definition 5)	
Level	**Quantile**
100% Max	8.50
99%	8.45
95%	7.65
90%	7.25
75% Q3	6.65
50% Median	5.40
25% Q1	4.20
10%	3.20
5%	2.70
1%	1.90
0% Min	1.30

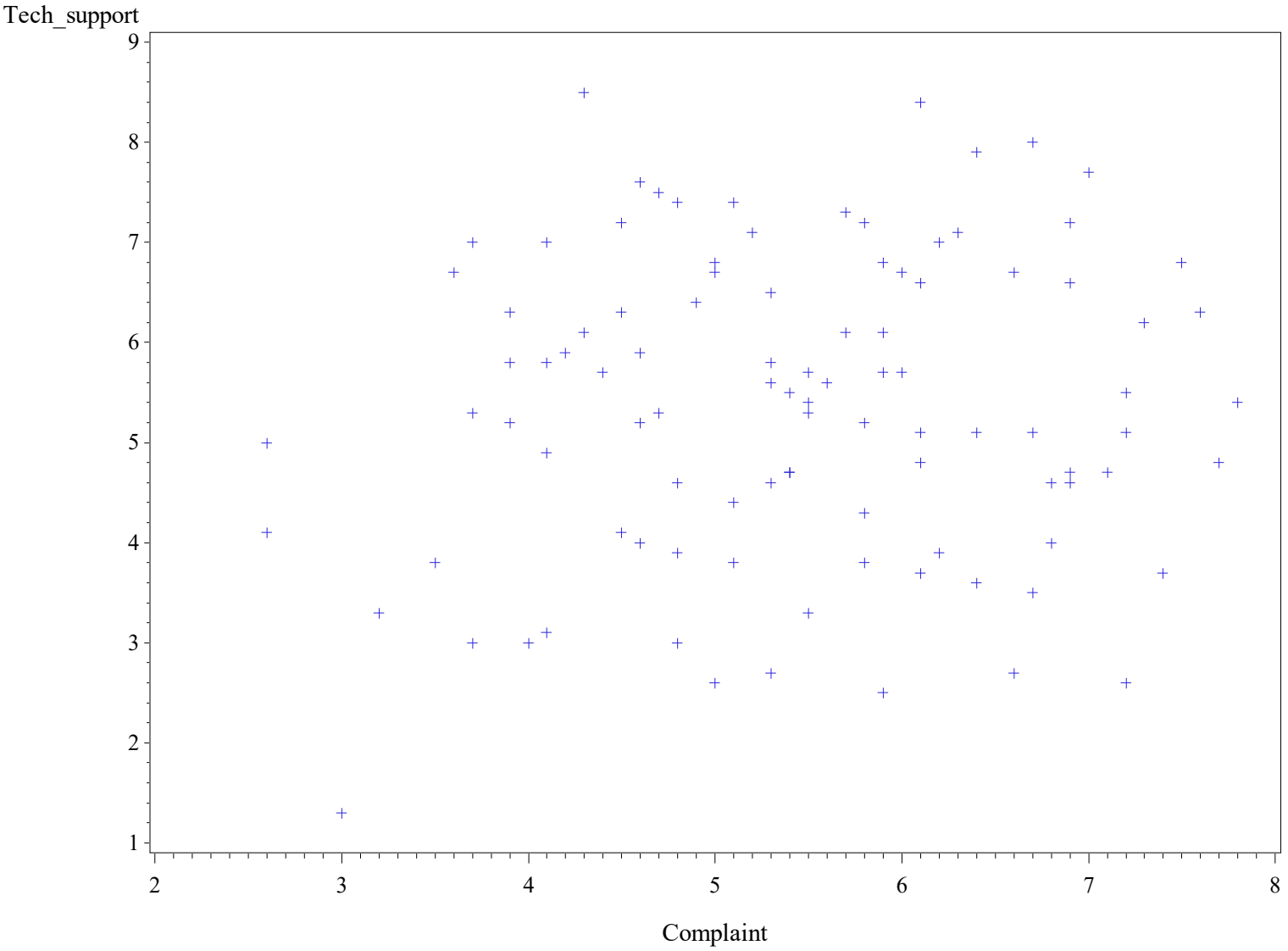
Extreme Observations			
Lowest		**Highest**	
Value	**Obs**	**Value**	**Obs**
1.3	87	7.7	90
2.5	1	7.9	67
2.6	97	8.0	88
2.6	24	8.4	61
2.7	73	8.5	31

The SAS System

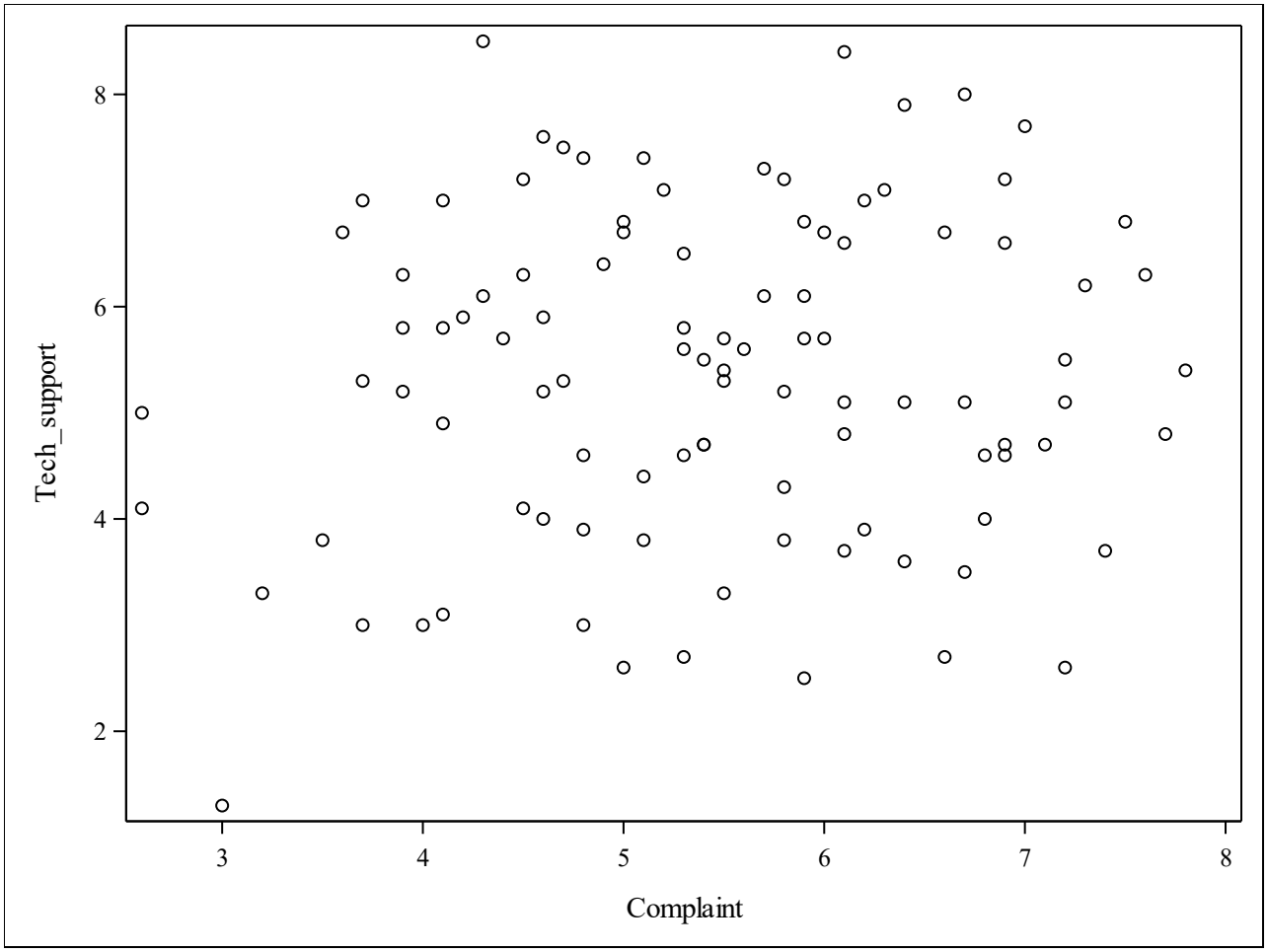
The UNIVARIATE Procedure



There is a positive linear relationship in the normal probability plot. The skew is nearly symmetrical and the data appears normally distributed. The whiskers of the box plot are similar in length, with the mean and median close to the center. The histogram appears to be multi-modal and validates a normal distribution. There are no outliers in the box plot.



The above and below scatter plots appears to be randomized with no noticeable pattern, suggesting homoscedasticity.


Tech_support
8
6
4
2
3
4
5
6
7
8
Complaint

The SAS System

The CORR Procedure

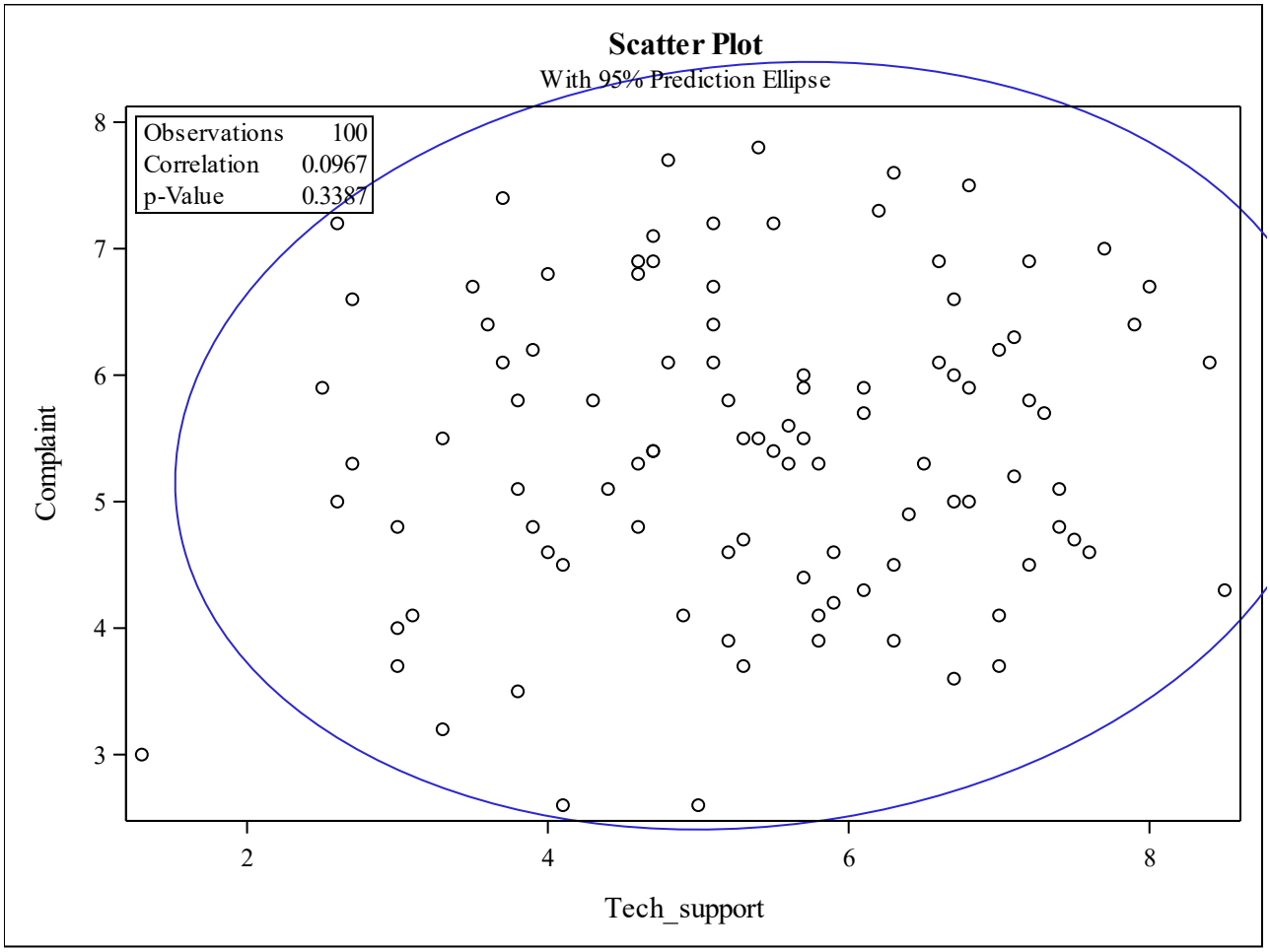
2 Variables:	Tech_support Complaint

Simple Statistics						
Variable	**N**	**Mean**	**Std Dev**	**Sum**	**Minimum**	**Maximum**
Tech_support	100	5.36500	1.53046	536.50000	1.30000	8.50000
Complaint	100	5.44200	1.20840	544.20000	2.60000	7.80000

Pearson Correlation Coefficients, N = 100 Prob > |r| under H0: Rho=0		
	Tech_support	**Complaint**
Tech_support	1.00000	0.09666 0.3387
Complaint	0.09666 0.3387	1.00000

The SAS System

The CORR Procedure



There are no values outside the ellipse in the above scatterplot. No bivariate outliers are observed.

The SAS System

The ANOVA Procedure

Class Level Information		
Class	**Levels**	**Values**
Customer_type	3	1 2 3

Number of Observations Read	100
Number of Observations Used	100

The SAS System

The ANOVA Procedure

Dependent Variable: Tech_support

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	4.3031870	2.1515935	0.92	0.4031
Error	97	227.5843130	2.3462300		
Corrected Total	99	231.8875000			

R-Square	Coeff Var	Root MSE	Tech_support Mean
0.018557	28.55062	1.531741	5.365000

Source	DF	Anova SS	Mean Square	F Value	Pr > F
Customer_type	2	4.30318696	2.15159348	0.92	0.4031

Testing: Means of the 3 customer types. Where 1= less than one year, 2 = between 1&5 years, 3= longer than 5 years. F(2,97)=0.92

H_0: $\mu_1 = \mu_2 = \mu_3$
H_1: Not all μ_j are equal.

The P-Value of 0.4031 > than $\alpha = 0.05$ Fail reject the null hypothesis. There is insufficient evidence that not all means are equal.

The SAS System

The ANOVA Procedure

Levene's Test for Homogeneity of Tech_support Variance ANOVA of Squared Deviations from Group Means					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Customer_type	2	10.4526	5.2263	0.73	0.4831
Error	97	691.5	7.1292		

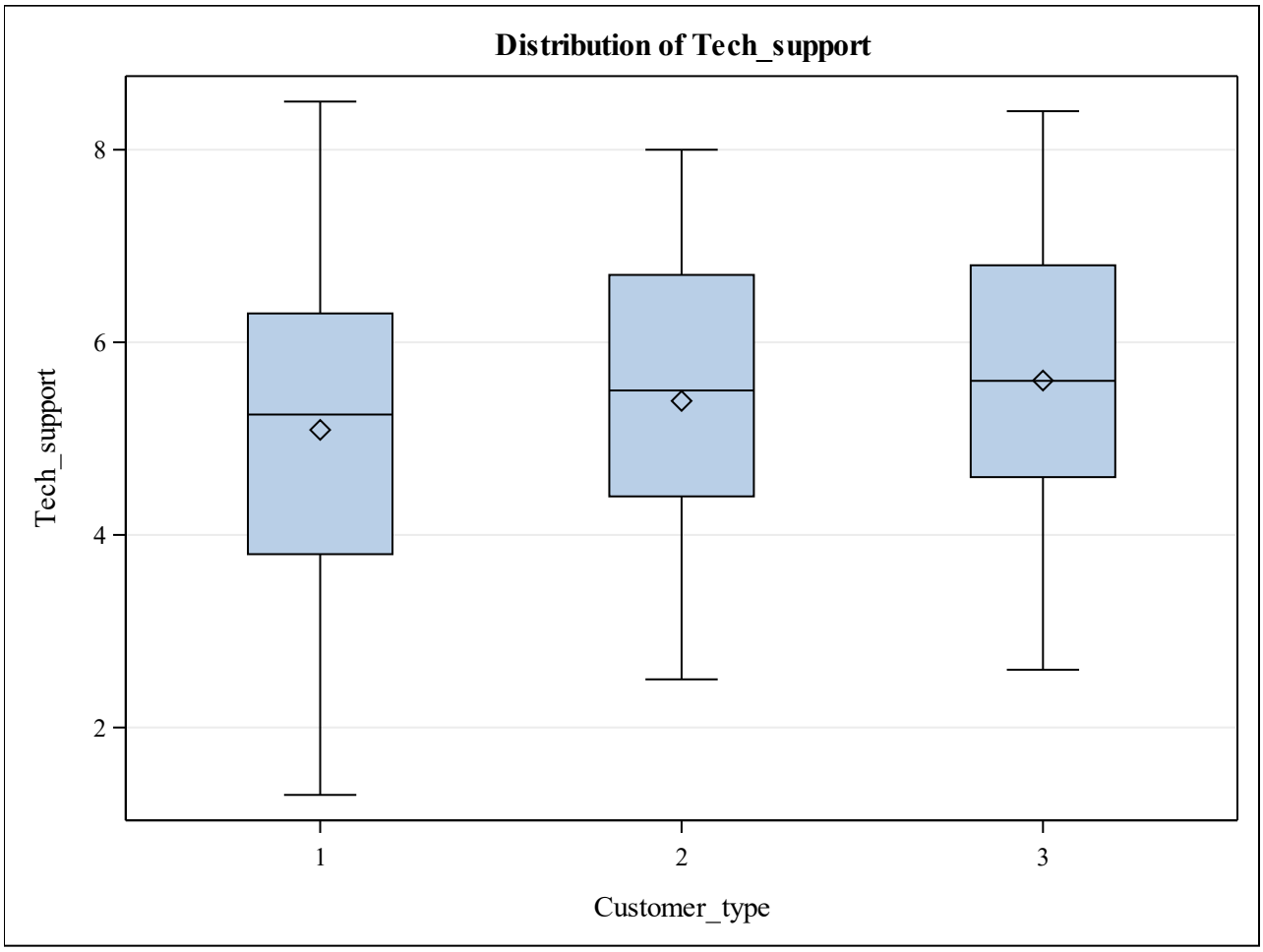
Levene's Test: Equal Variance of the 3 customer types. Where 1= less than one year, 2 = between 1&5 years, 3= longer than 5 years. F(2,97)=0.73

H0: $\sigma_1^2 = \sigma_2^2 = \sigma_3^2$
H1: Not all σ_j^2 are equal.

The P-Value of 0.4831 > than $\alpha = 0.05$ Fail reject the null hypothesis. There is insufficient evidence that not all variances are equal. The assumption of equal variances appears to be valid.

The SAS System

The ANOVA Procedure



Level of Customer_type	N	Tech_support Mean	Tech_support Std Dev
1	32	5.09062500	1.67473818
2	35	5.39142857	1.50555275
3	33	5.60303030	1.40945132

A boxplot of the data is used to approximate normality for each of the 3 Customer_types. The median is approximately centered, the whiskers are approximately equal, and the variability is similar on both sides of the IQR. There is a negative skew in number 1 and 2, with number 3 being nearly symmetric. Approximate normality is validated.

The SAS System

The REG Procedure
Model: MODEL1
Dependent Variable: Tech_support

Number of Observations Read	100
Number of Observations Used	100

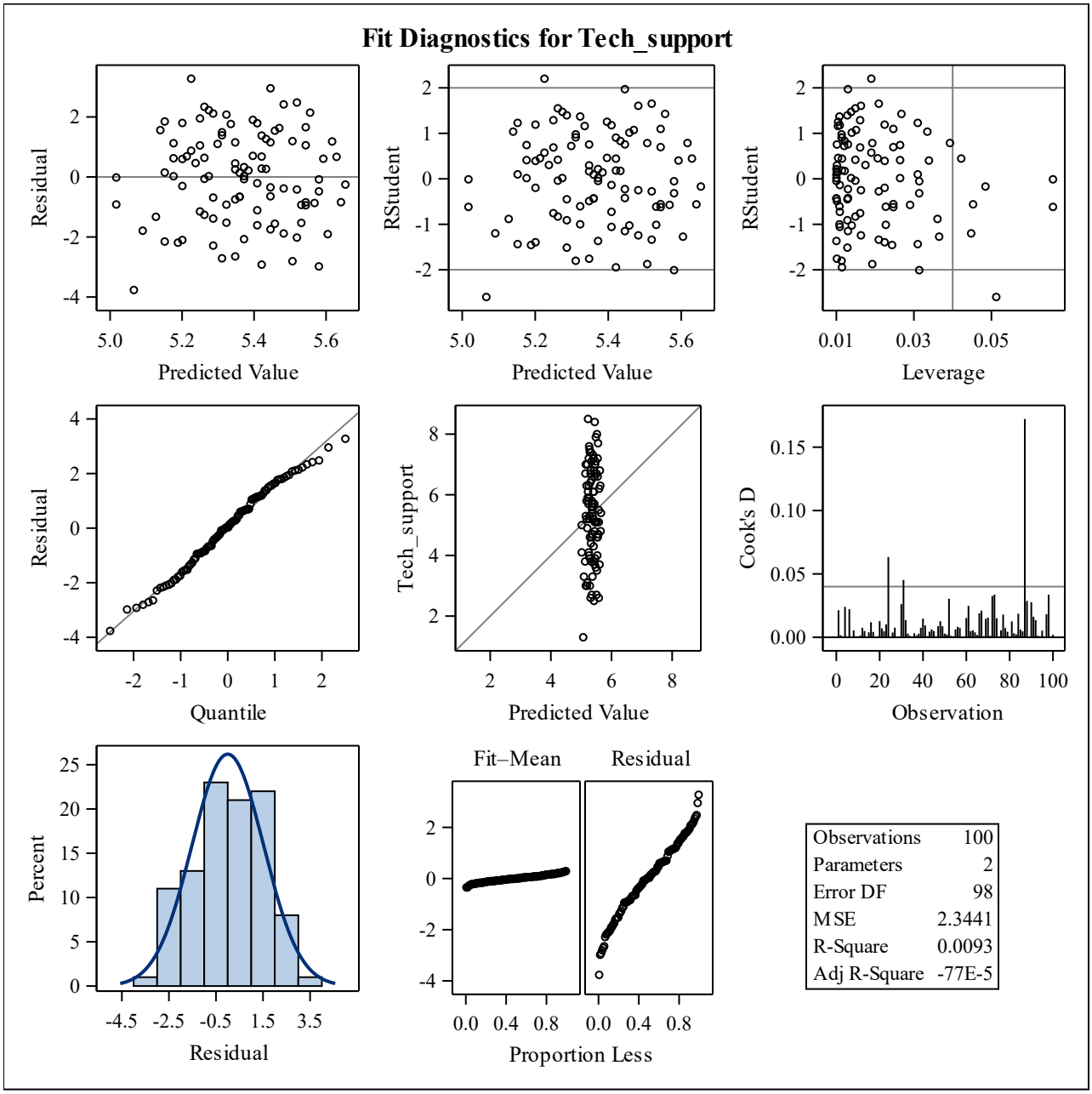
Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	1	2.16641	2.16641	0.92	0.3387
Error	98	229.72109	2.34409		
Corrected Total	99	231.88750			

Root MSE	1.53104	**R-Square**	0.0093
Dependent Mean	5.36500	**Adj R-Sq**	-0.0008
Coeff Var	28.53761		

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	1	4.69881	0.70969	6.62	<.0001
Complaint	1	0.12242	0.12734	0.96	0.3387

The SAS System

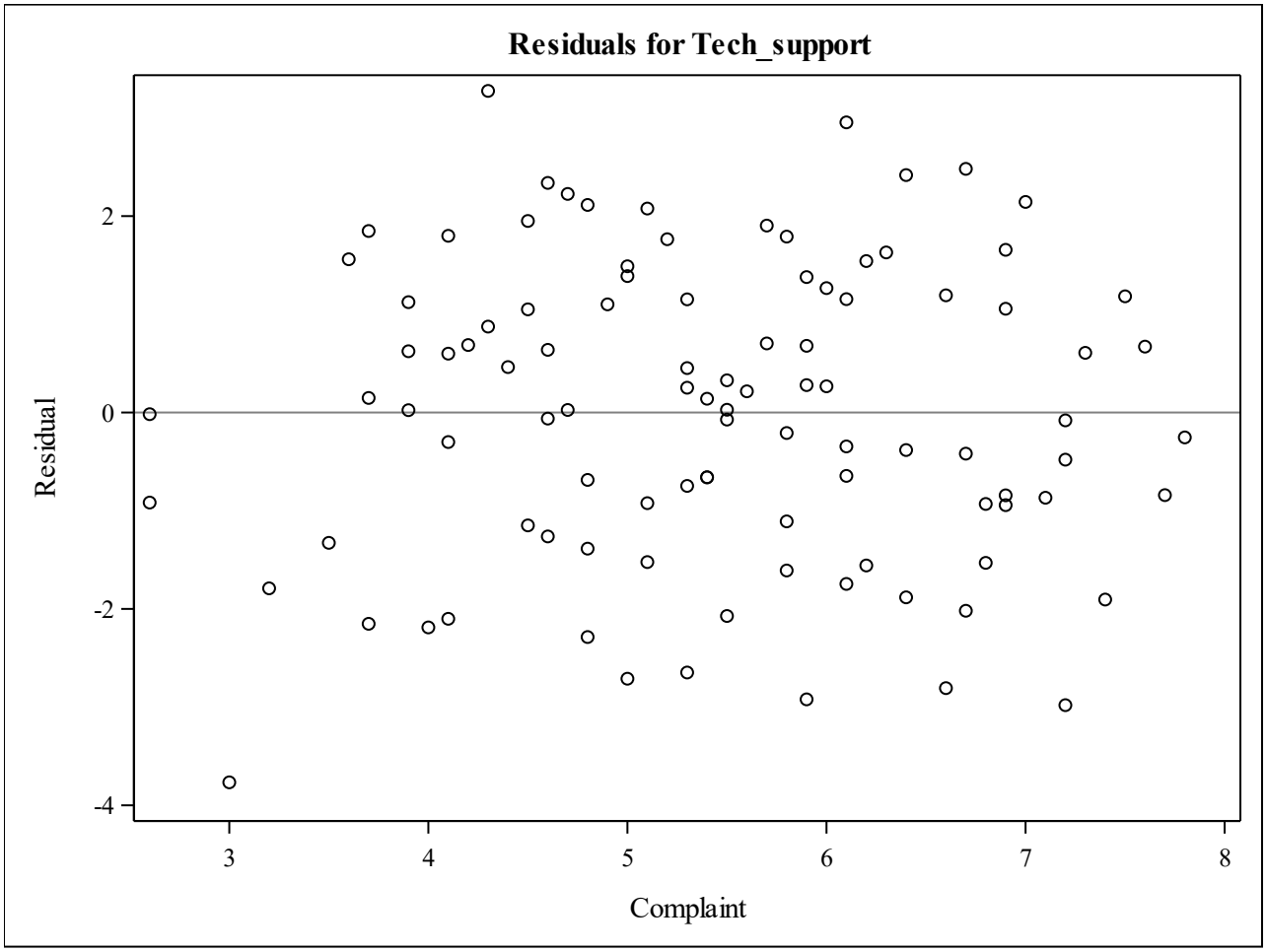
The REG Procedure
Model: MODEL1
Dependent Variable: Tech_support



From this histogram, it appears the residuals are normal, validating the assumption of normality. Residual versus predicted shows two potential outliers and appears to be randomized with no noticeable pattern, suggesting homoscedasticity. Residual versus Quantile shows a positive linear relationship with a slop close to 1, also validating an assumption of normal distribution.

The SAS System

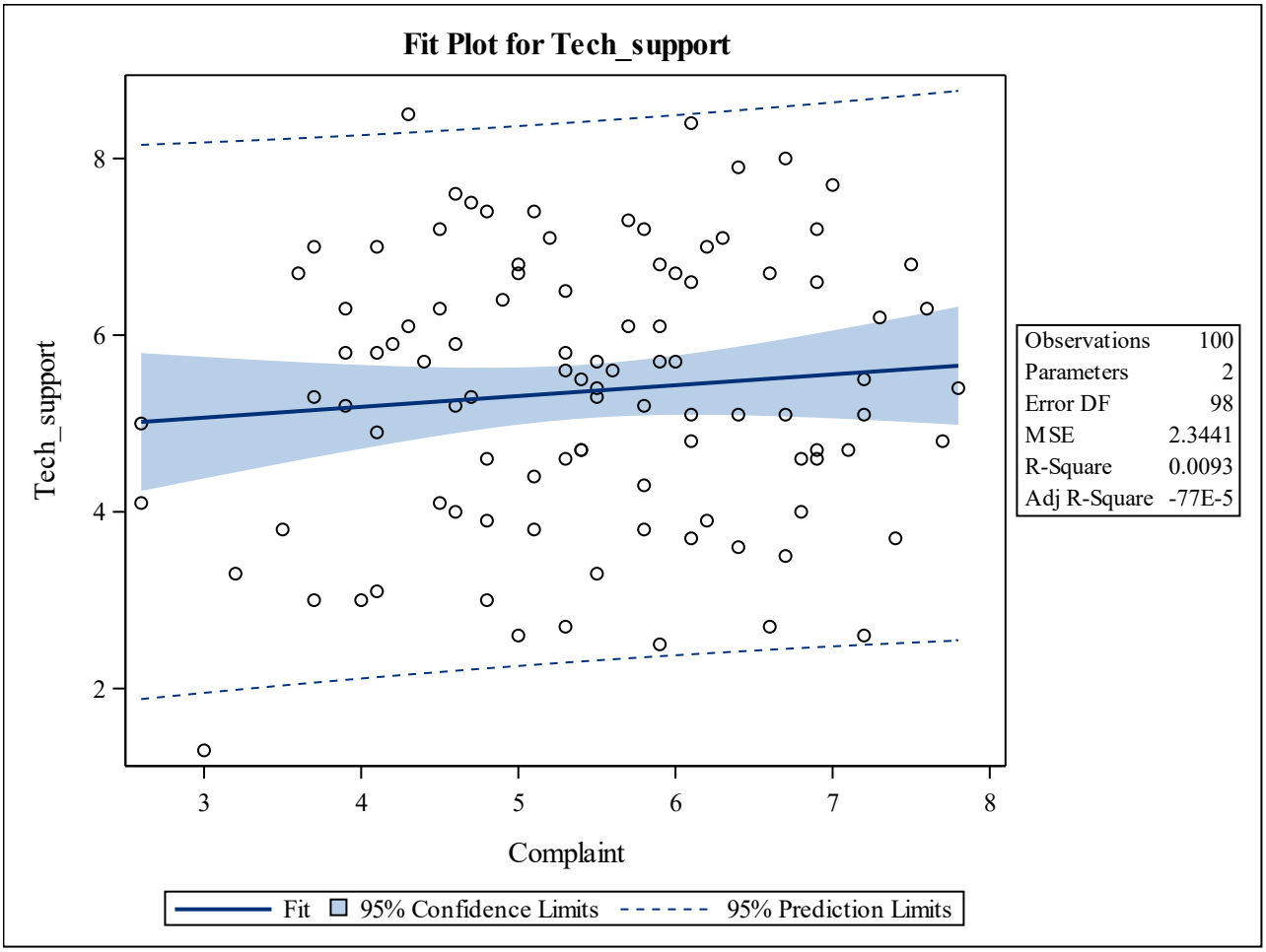
The REG Procedure
Model: MODEL1
Dependent Variable: Tech_support



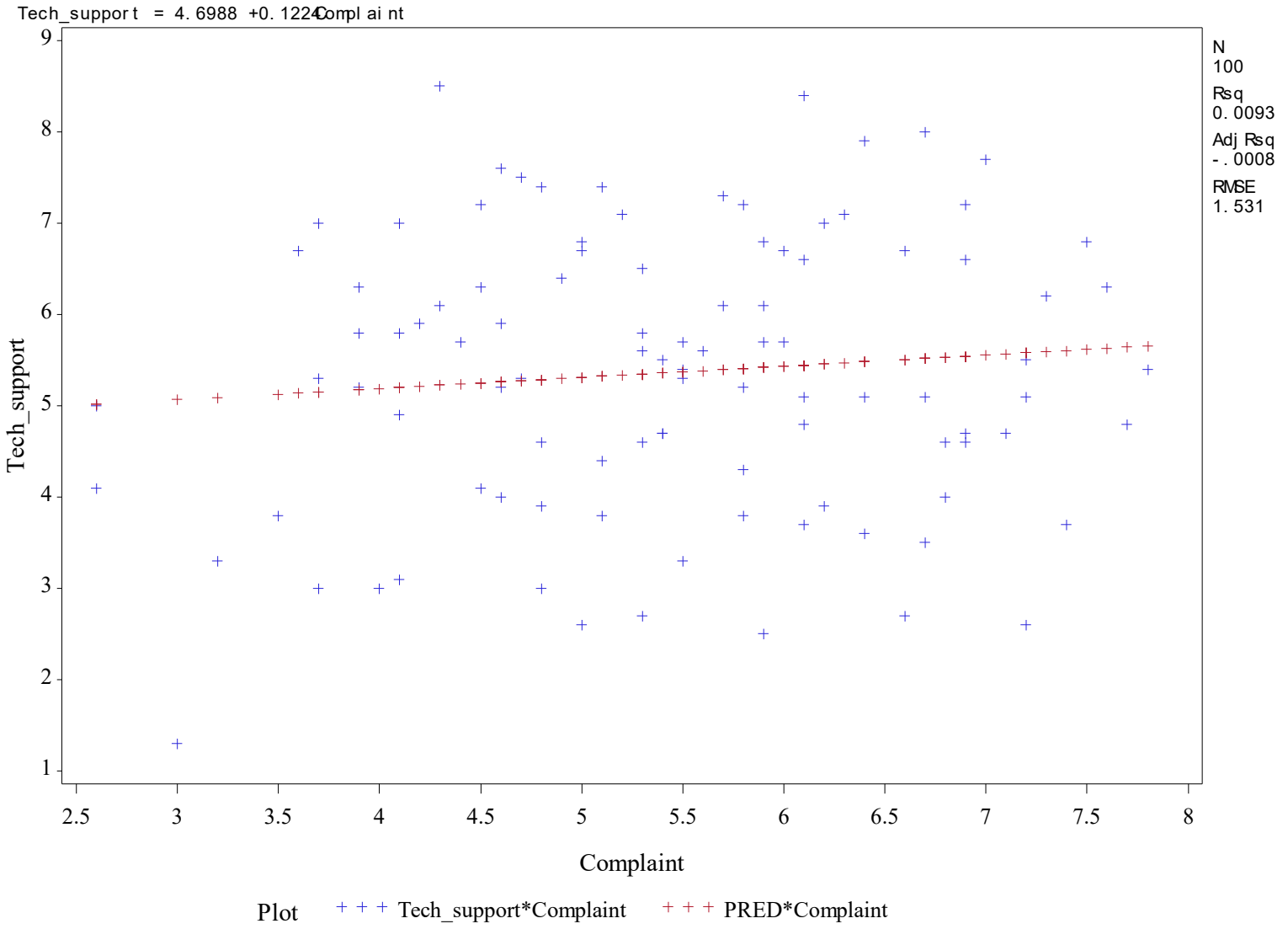
Residuals are normally distributed with about 68% of the residuals should be within 1 standard deviation of the mean of 0; 95% within 2 standard deviation, and about 99% within 3 standard deviations.

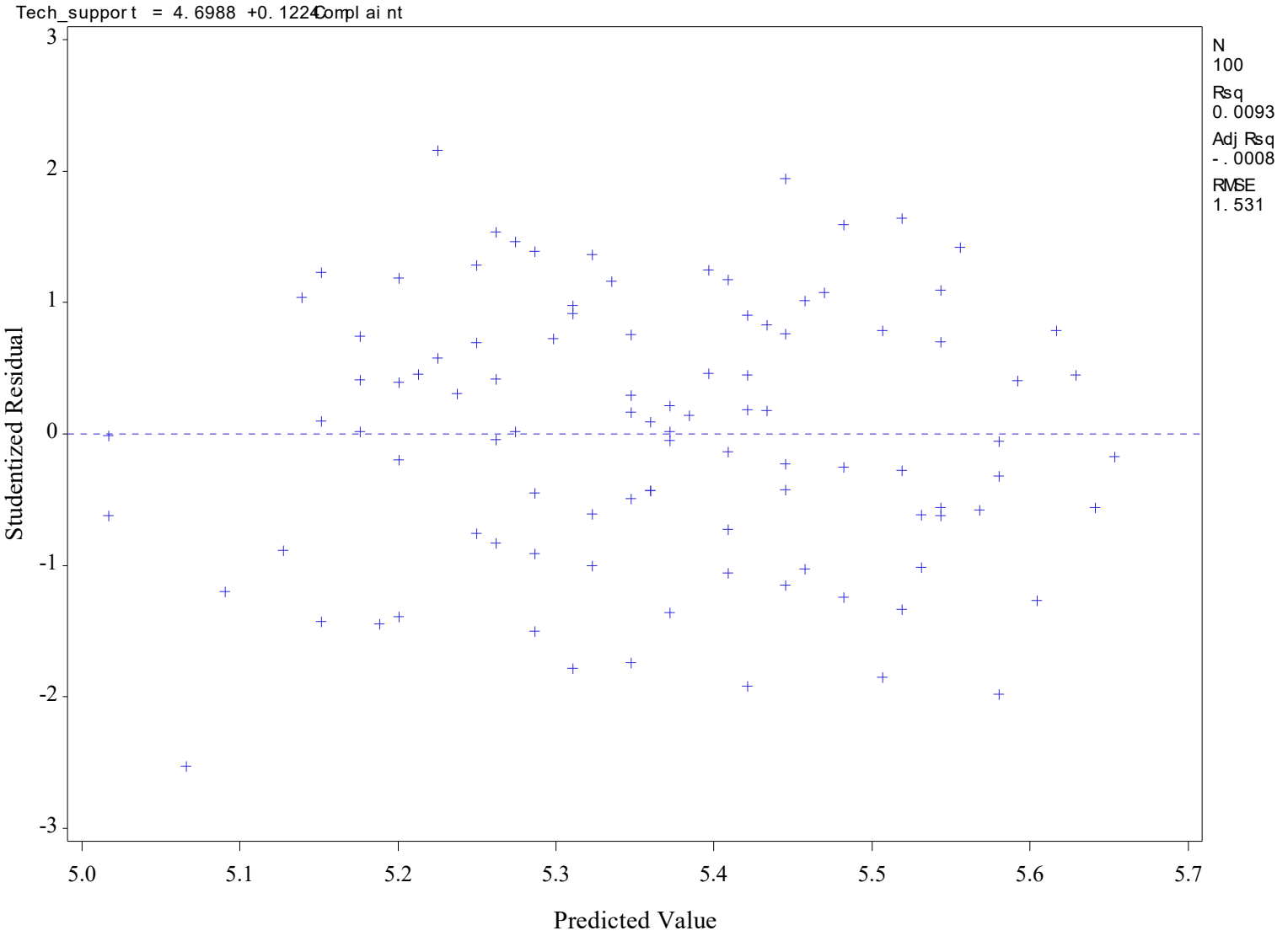
The SAS System

The REG Procedure
Model: MODEL1
Dependent Variable: Tech_support



The fitted plot shows normality with two outliers.


Tech_support = 4.6988 +0.1224Complaint
N
100
Rsq
0.0093
Adj Rsq
-.0008
RMSE
1.531
Tech_support
Complaint
Plot Tech_support*Complaint PRED*Complaint



Residuals are homoscedastic, there is no apparent pattern. The distribution of the error terms is assumed to be normal, about 68% of the residuals should be within 1 standard deviation of the mean, about 95% within 2 standard deviations, and 99% within 3.

The SAS System

The UNIVARIATE Procedure
Variable: stresid (Studentized Residual)

Moments			
N	100	**Sum Weights**	100
Mean	-0.0010047	**Sum Observations**	-0.1004723
Std Deviation	1.00541852	**Variance**	1.0108664
Skewness	-0.1789027	**Kurtosis**	-0.6170686
Uncorrected SS	100.075875	**Corrected SS**	100.075774
Coeff Variation	-100069.27	**Std Error Mean**	0.10054185

Basic Statistical Measures			
Location		**Variability**	
Mean	-0.00100	**Std Deviation**	1.00542
Median	0.01765	**Variance**	1.01087
Mode	-0.43316	**Range**	4.68494
		Interquartile Range	1.44908

Descriptive Values

Nearly symmetrical with a skewness value of -0.18.

A kurtosis value of -0.62 indicates that the distribution is slightly flat, but still a normal distribution.

The mean of -0.00 < 0.02 median also indicating a nearly symmetrical distribution.

Tests for Location: Mu0=0				
Test	**Statistic**		**p Value**	
Student's t	**t**	-0.00999	**Pr > |t|**	0.9920
Sign	**M**	2	**Pr >= |M|**	0.7644
Signed Rank	**S**	40	**Pr >= |S|**	0.8914

Location Test

H_0: $\mu=0$
H_1: $\mu\neq0$

t(99) = -0.01
A p-value 0.99 > α of .05. Fail to reject the null. The mean is statistically zero.

Tests for Normality				
Test	**Statistic**		**p Value**	
Shapiro-Wilk	**W**	0.987642	**Pr < W**	0.4822
Kolmogorov-Smirnov	**D**	0.064553	**Pr > D**	>0.1500
Cramer-von Mises	**W-Sq**	0.050063	**Pr > W-Sq**	>0.2500
Anderson-Darling	**A-Sq**	0.344713	**Pr > A-Sq**	>0.2500

Normality Test

H_0: Data is normal.
H_1: Data is not normal.

W=0.9876
A p-value of 0.48 > α of .05.
Fail to reject the null. The variable, Validating normal distribution.

The SAS System

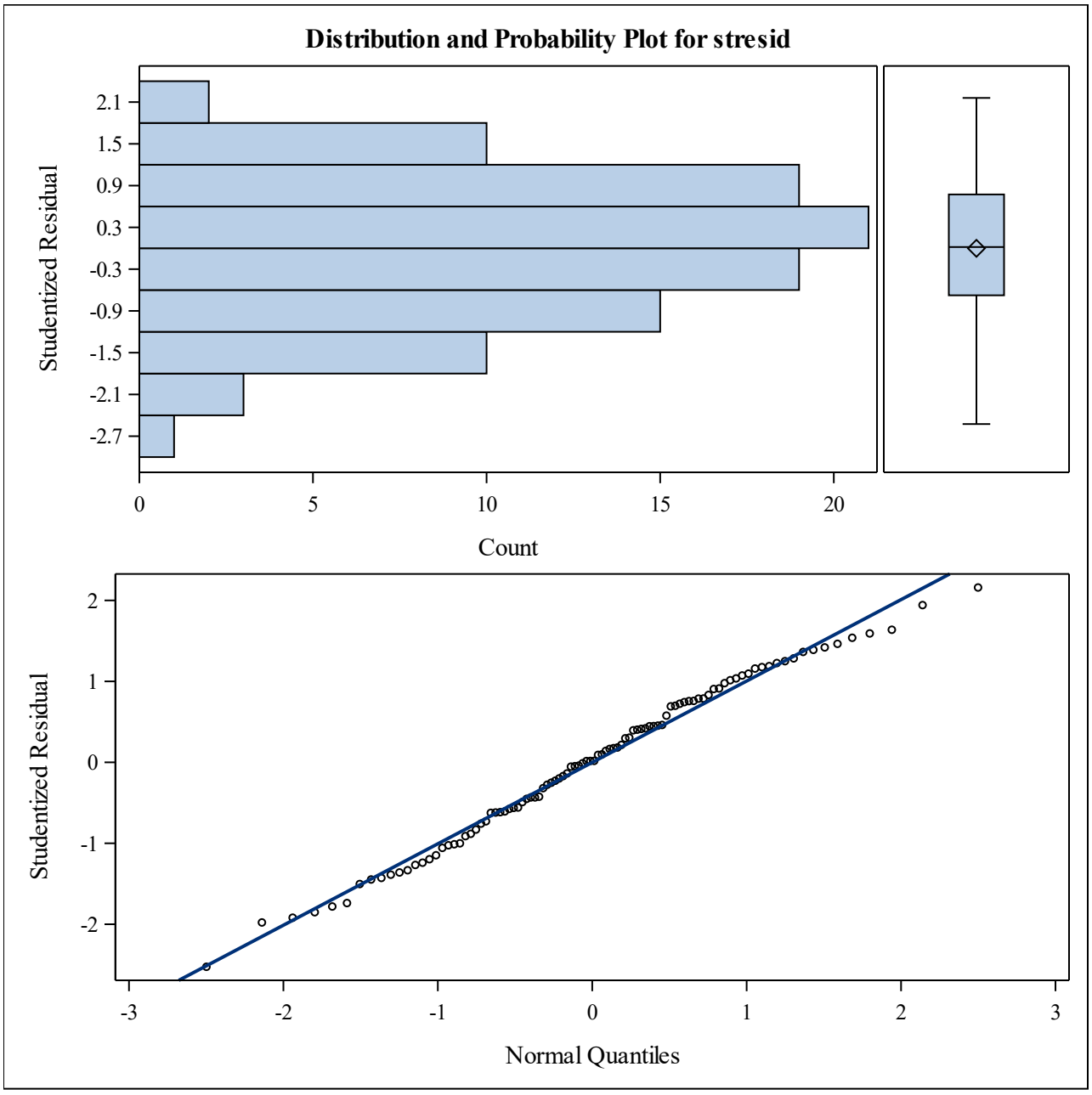
The UNIVARIATE Procedure
Variable: stresid (Studentized Residual)

Quantiles (Definition 5)	
Level	**Quantile**
100% Max	2.1595716
99%	2.0509654
95%	1.5012829
90%	1.2670113
75% Q3	0.7729814
50% Median	0.0176527
25% Q1	-0.6760975
10%	-1.3739895
5%	-1.7594387
1%	-2.2515815
0% Min	-2.5253637

Extreme Observations			
Lowest		**Highest**	
Value	**Obs**	**Value**	**Obs**
-2.52536	87	1.53862	77
-1.97780	24	1.59220	67
-1.91891	1	1.63771	88
-1.85116	73	1.94236	61
-1.78075	97	2.15957	31

The SAS System

The ANOVA Procedure



There is a positive linear relationship in the normal probability plot of studentized residuals. The skew is nearly symmetrical, and the data appears normally distributed. The whiskers of the box plot are similar in length, with the mean and median aligned. The histogram validates a normal distribution. There are no outliers in the box plot.