18

Continuous assignment 2

（a）Apply the log transformation to variable rating to create a new variable Lrating. Check Normality of rating and Lrating and briefly discuss the effect of the log transformation on the distribution.

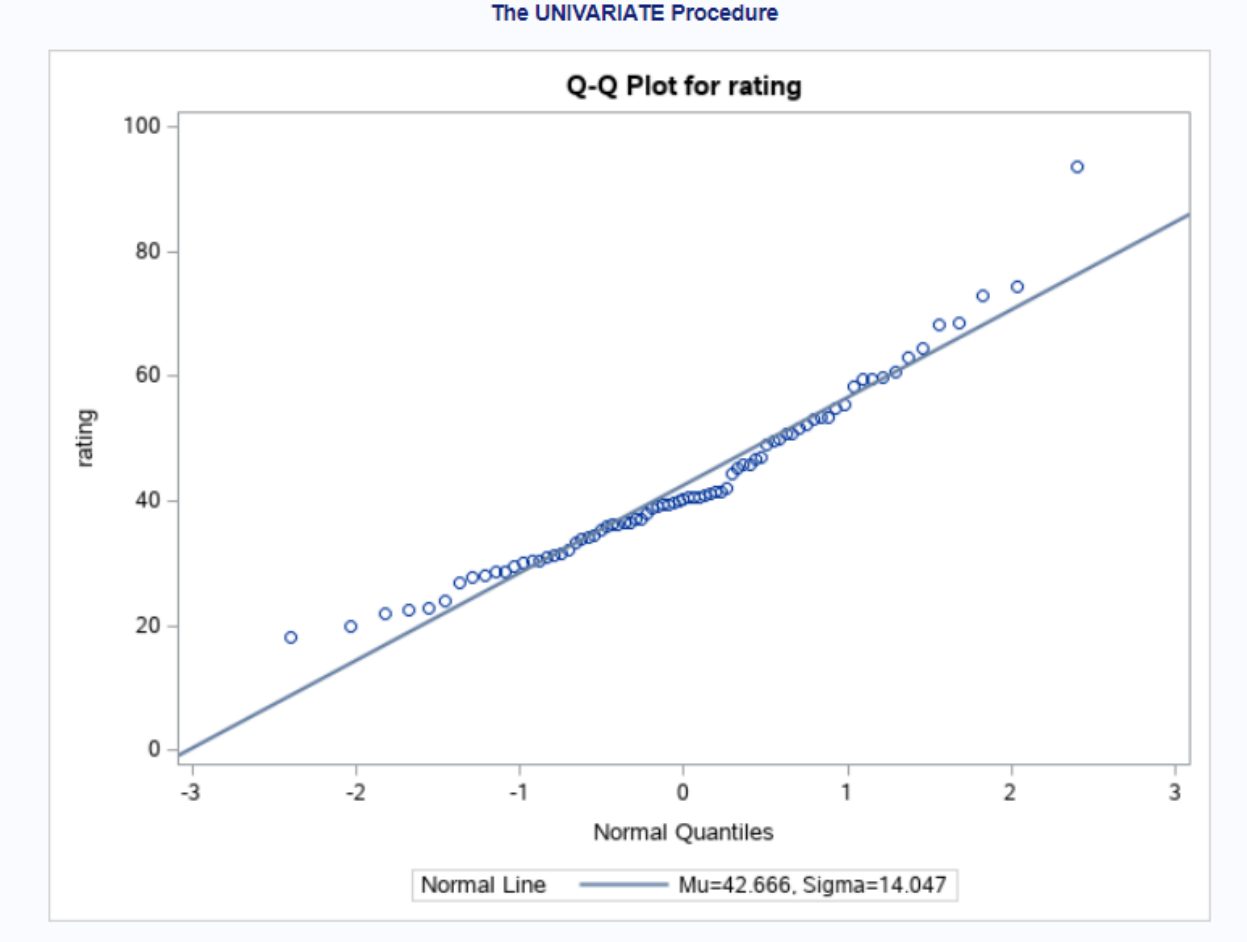
Basic description and tests for normality for ‘rating’:

The MEANS Procedure

Analysis Variable : rating									
N	Mean	Std Dev	Median	Minimum	Lower Quartile	Upper Quartile	Quartile Range	Skewness	Kurtosis
77	42.6657050	14.0472887	40.4002080	18.0428510	33.1740940	50.8283920	17.6542980	0.9102403	1.3187469

The UNIVARIATE Procedure
Variable: rating

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.953468	Pr < W	0.0067
Kolmogorov-Smirnov	D	0.13037	Pr > D	<0.0100
Cramer-von Mises	W-Sq	0.161088	Pr > W-Sq	0.0179
Anderson-Darling	A-Sq	0.892951	Pr > A-Sq	0.0223



It can be found from the descriptive statistics of 'rating': rating is not symmetrical, but right-skewed (skewness=0.9102403) and leptokurtic (kurtosis=1.3187469).

Typical values for 'rating' are: the median is 40.4, the mean is 42.67, the minimum is 18.04, and the maximum is 40.40.

It is not difficult to see from the Q-Q graph that the distribution of rating is not a normal distribution; the P-values of the four tests used in the normality test are all less than 0.05, which is also a strong evidence that the distribution of rating is not a normal distribution.

request and interpret the histogram as well.

Then, apply the log transformation to variable rating to create a new variable Lrating.

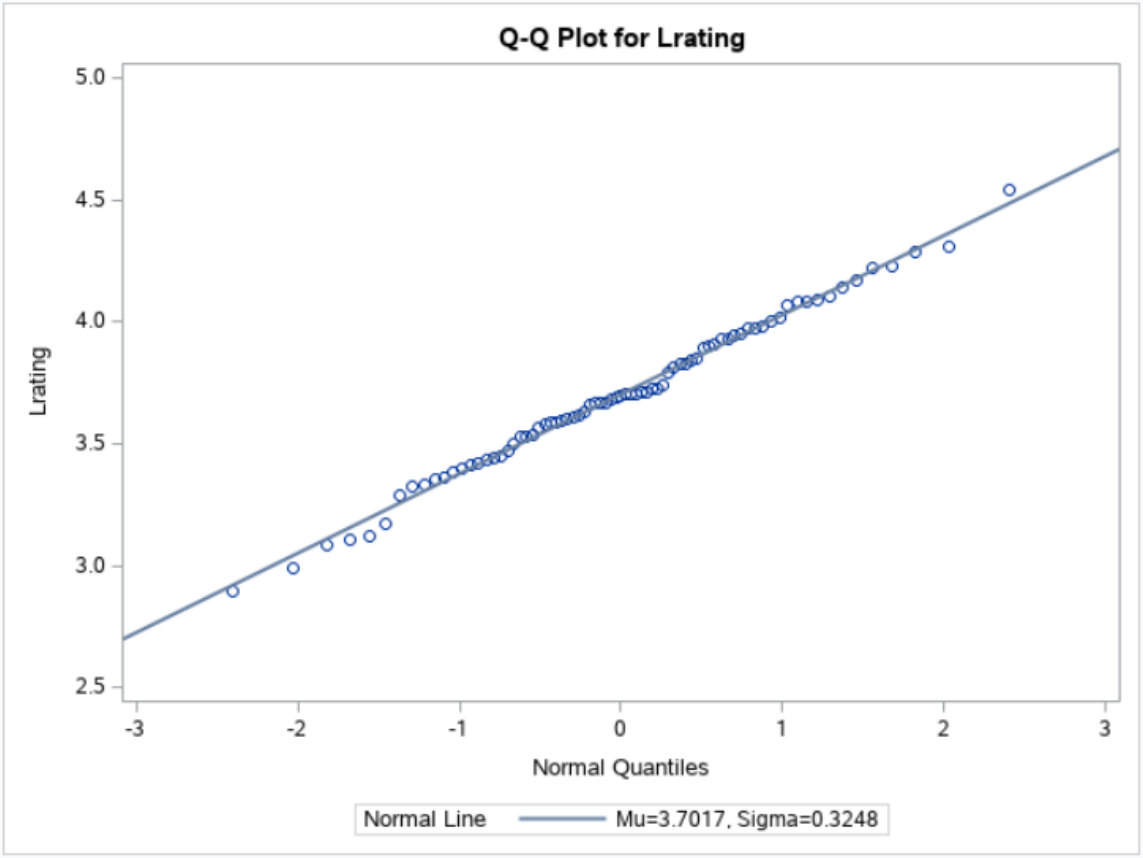
The MEANS Procedure

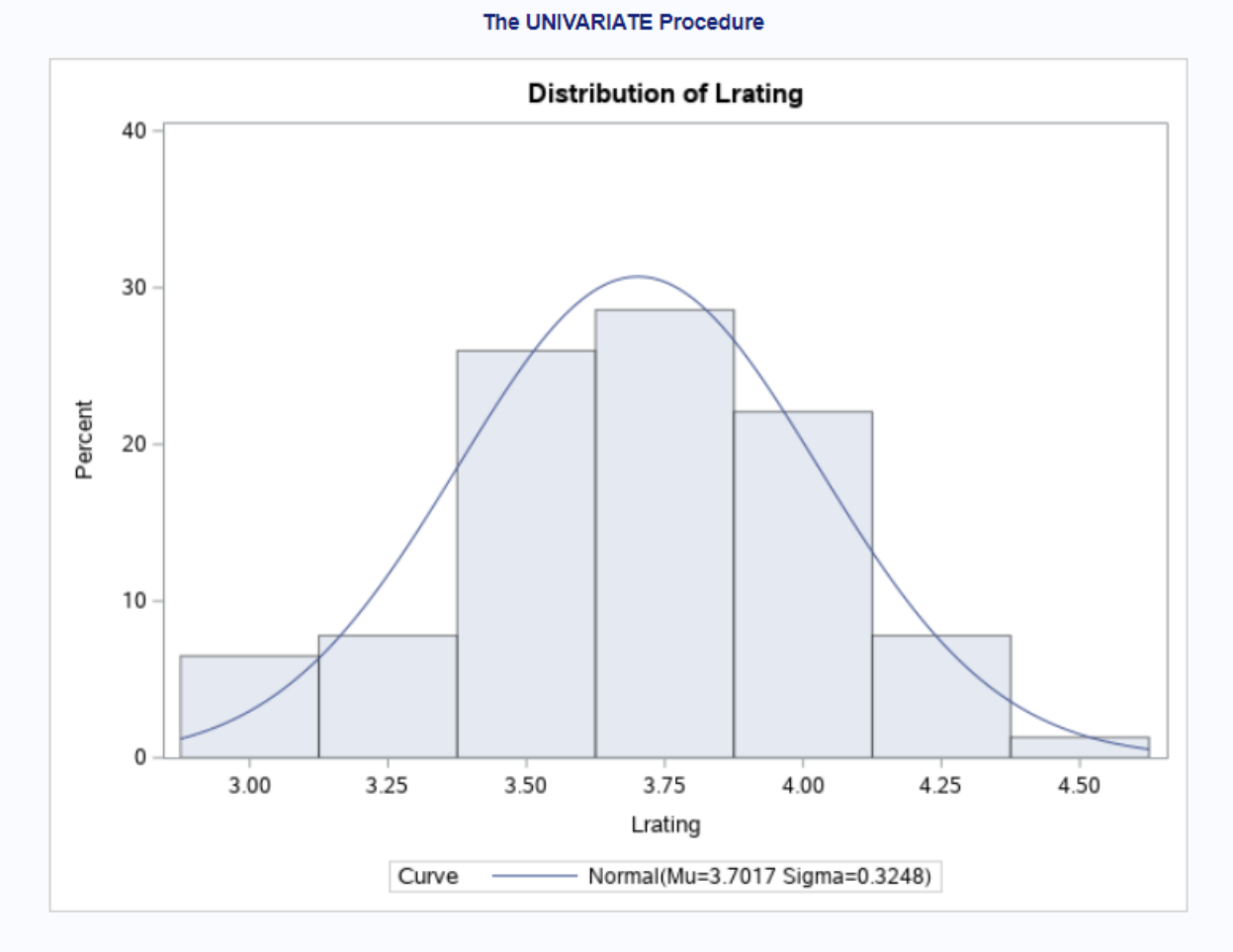
Analysis Variable : Lrating									
N	Mean	Std Dev	Median	Minimum	Lower Quartile	Upper Quartile	Quartile Range	Skewness	Kurtosis
77	3.7017049	0.3248353	3.6988349	2.8927495	3.5017693	3.9284551	0.4266858	-0.0648320	0.0076681

The UNIVARIATE Procedure
Variable: Lrating

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.994291	Pr < W	0.9829
Kolmogorov-Smirnov	D	0.067864	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.030962	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.182532	Pr > A-Sq	>0.2500

The UNIVARIATE Procedure





After performing Log transformation, we obtain a new variable Lrating, and by observing the Histogram of Lrating, we can think that Lrating is basically symmetrically distributed. Through the descriptive statistics table, we can also find that the skewness and kurtosis of Lrating are -0.0648320 and 0.0076681.

Its typical value is 3.7 which is the mean value of Lrating.

The normality test showed p-values greater than 0.05 in all 4 tests, i.e. H0 could not be rejected. The Q-Q plot also shows a curve that approximates a straight line. From this, we can think that Lrating can be a normal distribution. The assumptions for establishing the Pearson correlation coefficient are satisfied.

you need normality for the other variables as well.

(b) Obtain a Pearson correlation matrix relating variables Lrating, sugars, fiber and sodium, and comment briefly on these correlations.

The CORR Procedure

4 Variables:	Lrating sugars fiber sodium

Pearson Correlation Coefficients
Prob > |r| under H0: Rho=0
Number of Observations

	Lrating	sugars	fiber	sodium
Lrating	1.00000 77	-0.77076 <.0001 76	0.53381 <.0001 77	-0.36990 0.0009 77
sugars	-0.77076 <.0001 76	1.00000 76	-0.13876 0.2319 76	0.05887 0.6135 76
fiber	0.53381 <.0001 77	-0.13876 0.2319 76	1.00000 77	-0.07068 0.5413 77
sodium	-0.36990 0.0009 77	0.05887 0.6135 76	-0.07068 0.5413 77	1.00000 77

Pearson Correlation Statistics (Fisher's z Transformation)								
Variable	With Variable	N	Sample Correlation	Fisher's z	Bias Adjustment	Correlation Estimate	95% Confidence Limits	p Value for H0:Rho=0
Lrating	sugars	76	-0.77076	-1.02221	-0.00514	-0.76867	-0.847290 -0.657089	<.0001
Lrating	fiber	77	0.53381	0.59546	0.00351	0.53129	0.348825 0.674954	<.0001
Lrating	sodium	77	-0.36990	-0.38831	-0.00243	-0.36780	-0.546738 -0.156732	0.0008
sugars	fiber	76	-0.13876	-0.13966	-0.0009251	-0.13785	-0.352357 0.090413	0.2328
sugars	sodium	76	0.05887	0.05893	0.0003924	0.05848	-0.169211 0.280236	0.6146
fiber	sodium	77	-0.07068	-0.07079	-0.0004650	-0.07021	-0.289636 0.156223	0.5425

Spearman Correlation Statistics (Fisher's z Transformation)								
Variable	With Variable	N	Sample Correlation	Fisher's z	Bias Adjustment	Correlation Estimate	95% Confidence Limits	p Value for H0:Rho=0
Lrating	sugars	76	-0.81059	-1.12874	-0.00540	-0.80873	-0.874696 -0.713333	<.0001
Lrating	fiber	77	0.48938	0.53524	0.00322	0.48693	0.295136 0.640998	<.0001
Lrating	sodium	77	-0.24257	-0.24750	-0.00160	-0.24106	-0.441220 -0.018060	0.0332
sugars	fiber	76	-0.09940	-0.09973	-0.0006627	-0.09874	-0.317140 0.129596	0.3942
sugars	sodium	76	-0.05378	-0.05383	-0.0003585	-0.05342	-0.275558 0.174131	0.6456
fiber	sodium	77	-0.16859	-0.17022	-0.00111	-0.16752	-0.377337 0.058664	0.1431

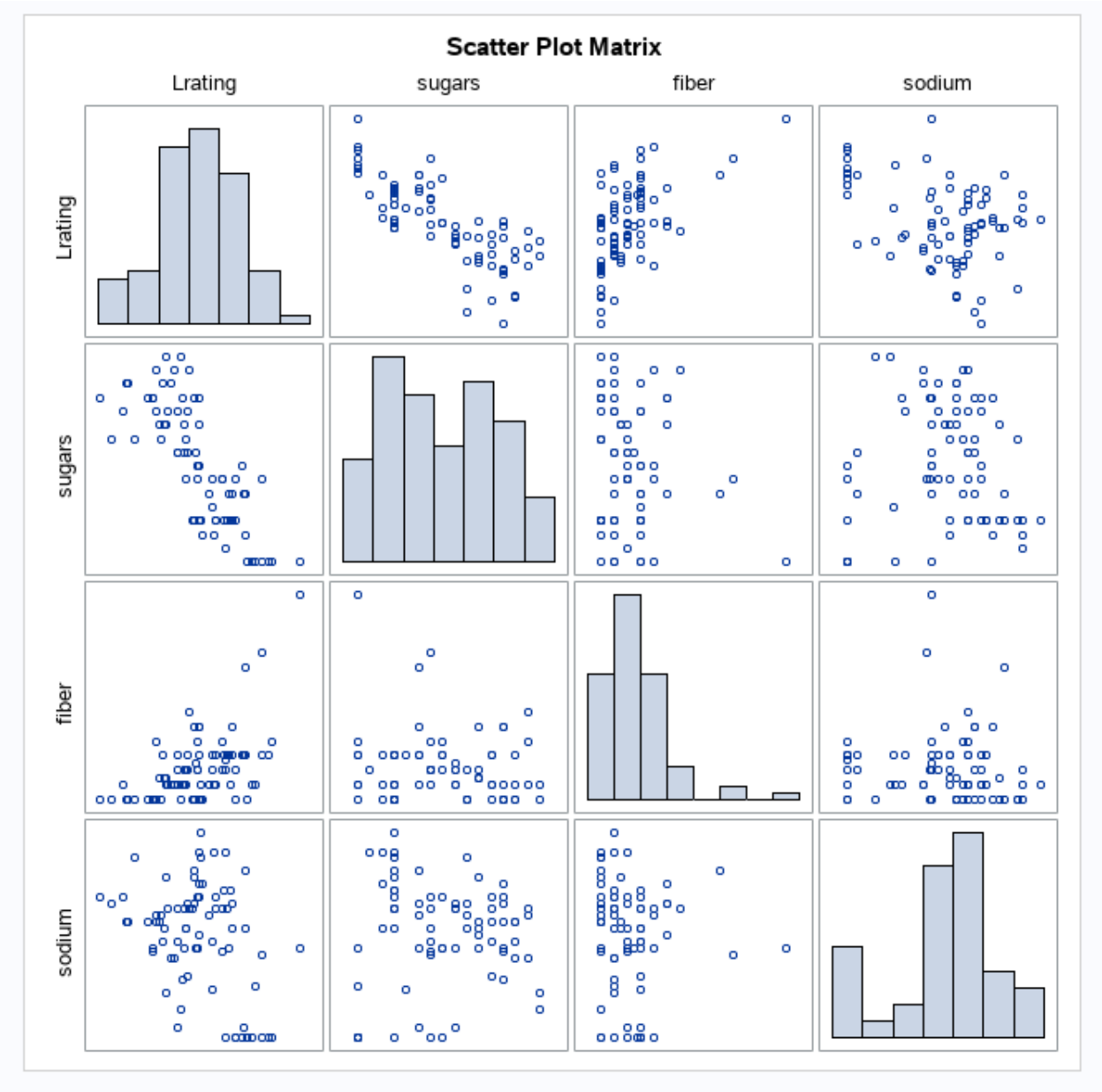
Hypothesis tests are based on

1. H0: r=0
2. H1: r≠0 and α=0.05

Through the values of r and p-value, it is not difficult to find that Lrating and sugars, fiber and sodium have significant correlations before three variables. The correlations between sugar, fiber and sodium are all significantly higher than 0.05, so H0 cannot be rejected, and there is not enough evidence to prove that there is a correlation between the three.

From the output of Firsher's z-transform, the 95% confidence limits show the smallest negative correlation error between Lrating and sugar. We can say with 95% confidence that the overall correlation coefficient between these two variables is between -0.847290 and -0.657089, indicating a large effect. For pairs of variables for Lrating and Fiber, the confidence limits indicate a positive correlation, with moderate to large effects. For the Lrating and Sodium pair, the negative correlation shows a wide range of effect sizes from a large -0.546738 to a small -0.156732.

(c) Obtain a scatterplot matrix relating Lrating, sugars, fiber and sodium, and briefly discuss the resulting relationships. Based on your scatterplots and results from part (b), which variable would you recommend as the single explanatory variable in a simple linear regression model for Lrating.



From the scatter plot, there is a linear relationship between Lrating and sugar; and the Pearson correlation coefficient given in the table also points out that the p-value is less than 0.0001, which means that the correlation is statistically significant at the 5% level .

We can think that the sugar variable should be used in the regression model because sugar has a relatively significant linear correlation and Lrating relative to fiber and sodium.

From the scatterplots matrix, there are non-linear patterns between Lrating and fiber and sodium, so the spearman correlation coefficient will be better.

Spearman's correlation coefficient showed that Lrating and Spearman Rho for fiber, Lrating and sodium were significant at the 5% level. So there is a relationship between Lrating and fiber and a relationship between Lrating and sodium.

(d)Fit a simple linear regression model relating Lrating to the variable you have identified in part (c), with Lrating as the dependent variable. Interpret the model equation. Obtain and discuss fit diagnostics. Are there any observations that would require closer inspection? Explain briefly.

We choose the variable Sugars.

The REG Procedure
Model: MODEL1
Dependent Variable: Lrating

Number of Observations Read	77
Number of Observations Used	76
Number of Observations with Missing Values	1

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	4.73318	4.73318	108.30	<.0001
Error	74	3.23409	0.04370		
Corrected Total	75	7.96727			

Root MSE	0.20905	R-Square	0.5941
Dependent Mean	3.69872	Adj R-Sq	0.5886
Coeff Var	5.65209		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	4.10184	0.04556	90.04	<.0001	4.01106	4.19262
sugars	1	-0.05737	0.00551	-10.41	<.0001	-0.06836	-0.04639

According to the table, we can get:

$$Lrating = 4.101 - 0.057 * Sugars$$

The p-value of the intercept is less than 0.0001, and the p-value of the slope is also less than 0.0001, which shows that the null hypothesis of the intercept and the slope of 0 cannot be established.

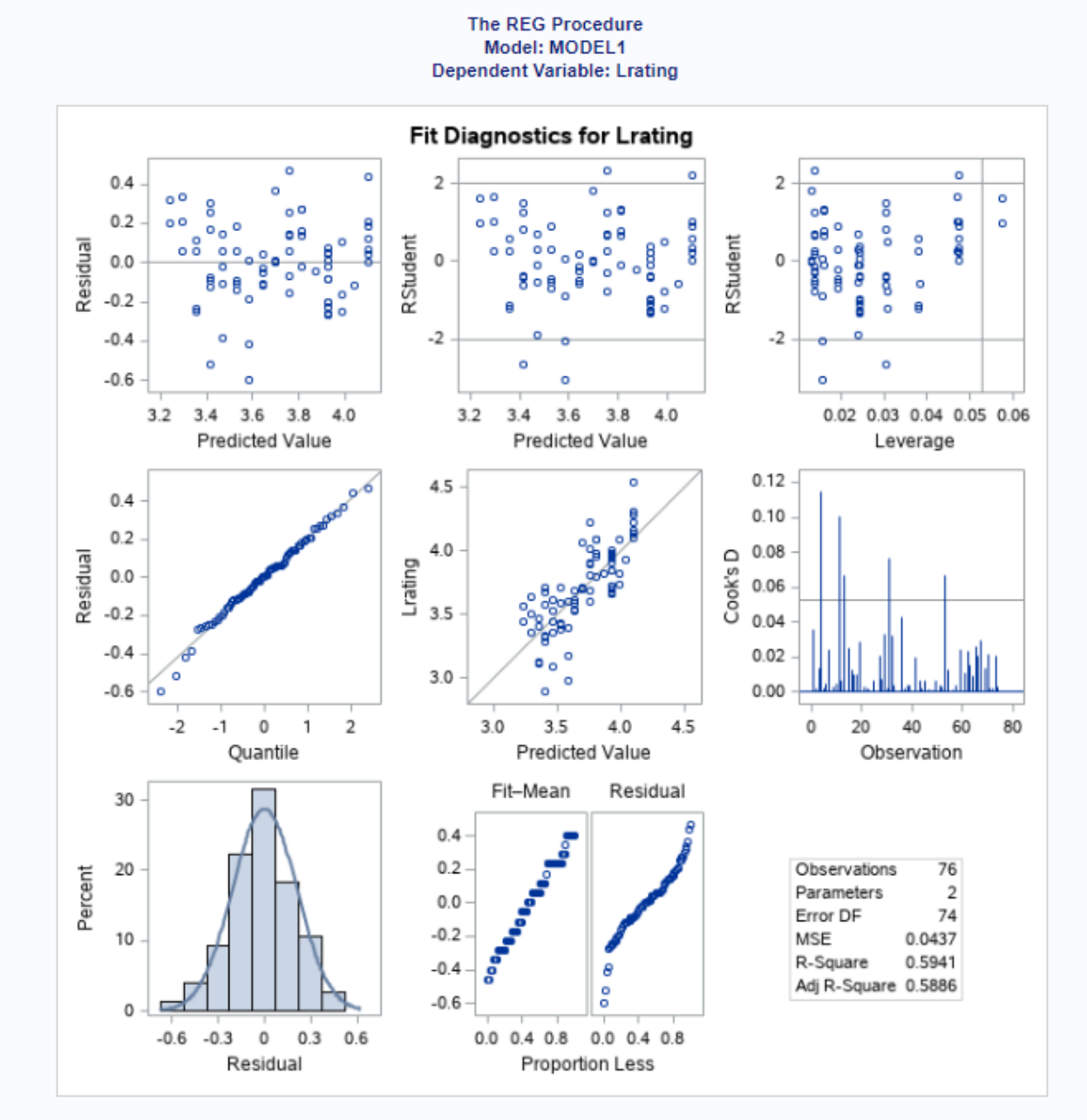
An intercept of 4.101 and a slope of -0.057 are statistically significant at the 5% level. The value of F in the ANOVA table is 108.3, and the value of p is also less than 0.0001. These data can prove the establishment of this significance.

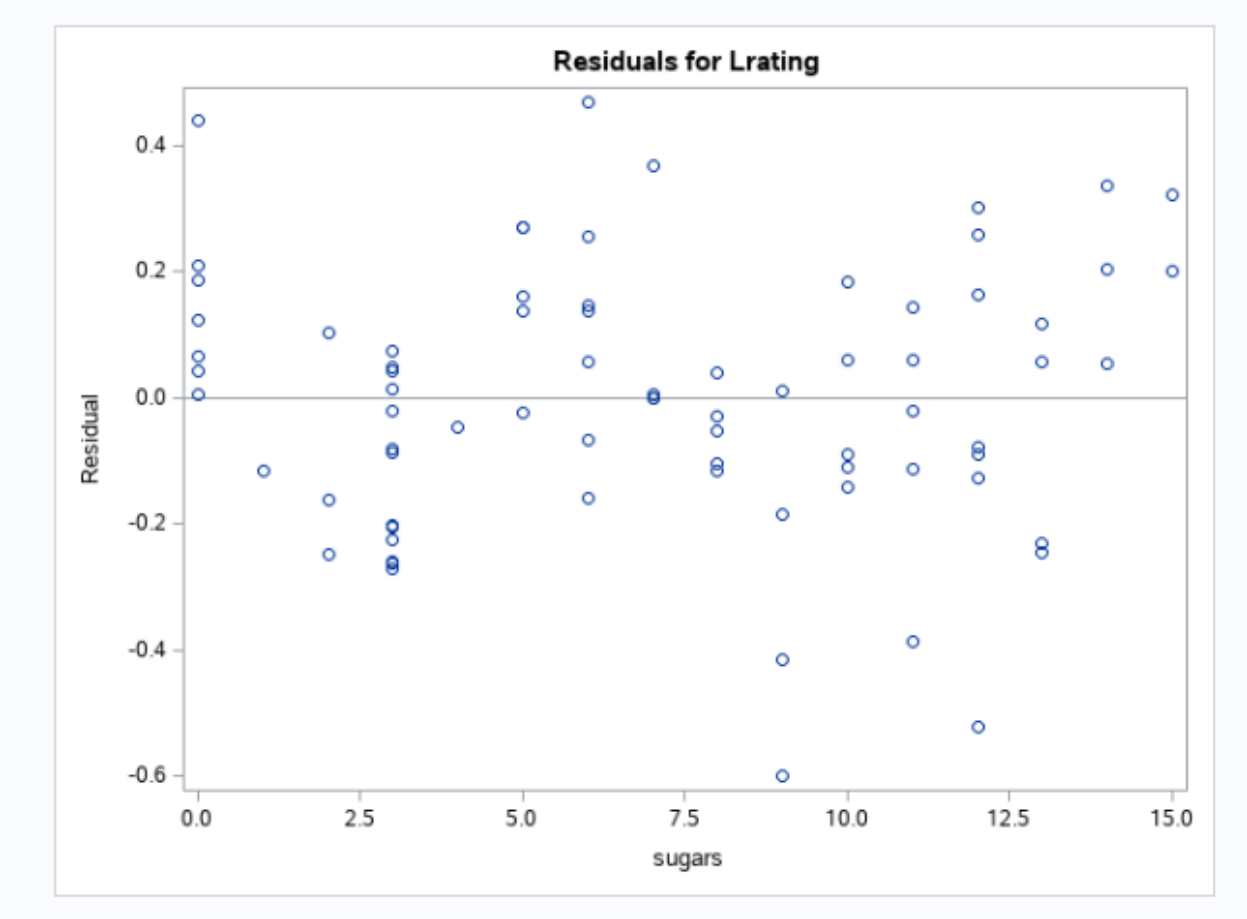
model is statistically significant; significant relationship b/w sugars and lrating

A statistically significant relationship between Lrating and Sugars is established, and this model can be used to predict the Population relationship between Sugars and Lrating.

Statistically speaking, the Lrating value of cereal products with 0 sugar content is about 4.101, which can correspond to the initial rating score of 60. We can say with 95% confidence that the expected population Lrating without sugar should have a score between 4.011 and 4.192, which corresponds to a raw score between 55 and 66.

How about interpretation of the slope?

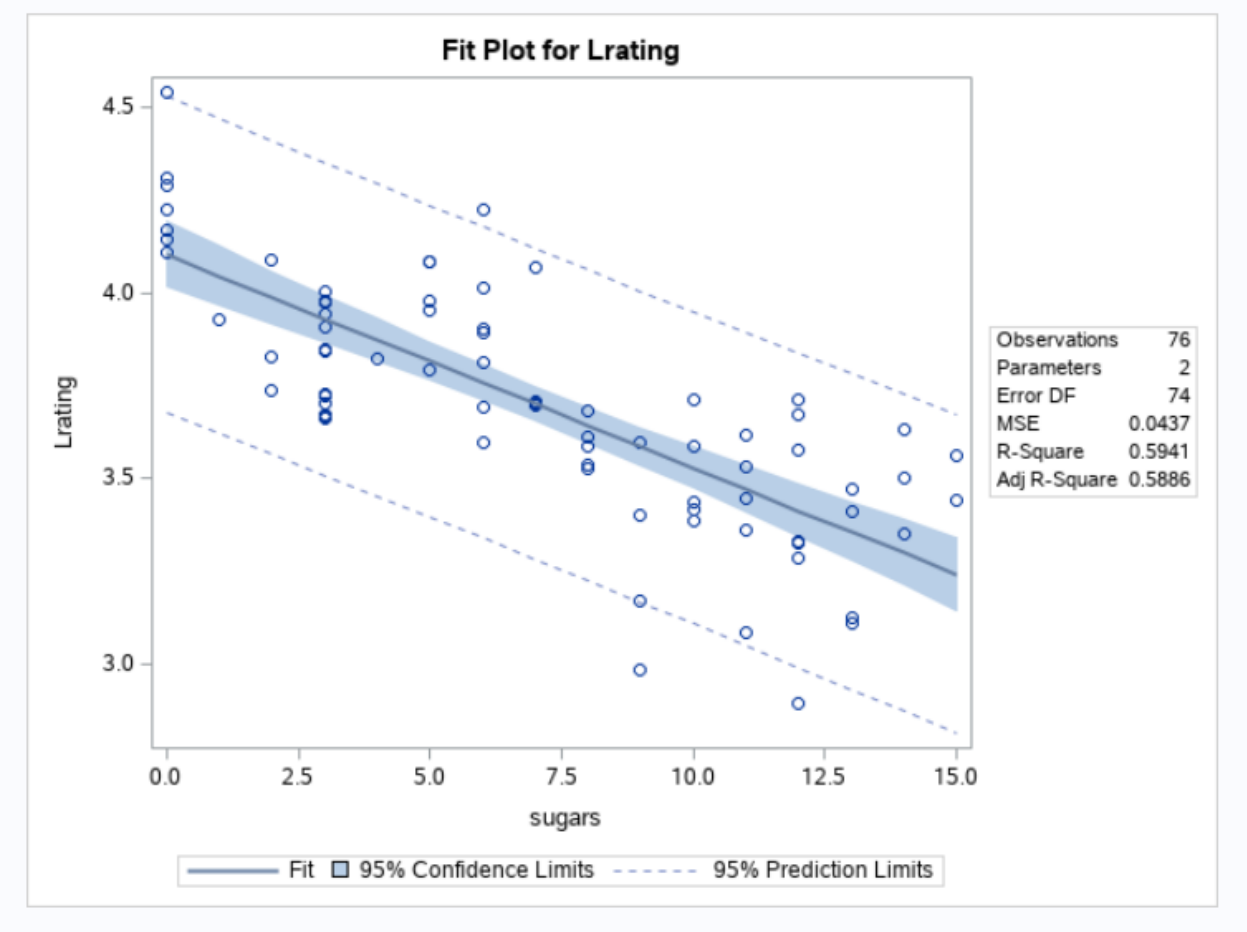




The residual value and the predicted value satisfy the linearity and error independence. According to the residual of the predicted value, the residual and predicted value do not show an obvious pattern or direction.

However, homoscedasticity behaves differently. The homoscedasticity will move from left to right with the parameter value, and the vertical distance from the residual value to the 0 line will first increase from small to large, and then from large to small. This behaviour is contrary to the "constant variance" of the linear regression model. Further investigation should be carried out.

The Q-Q plot shows that the residual predicted values do not have a curve pattern, which satisfies normality. The histogram also shows that the distribution of the residuals is close to the normal distribution.



The above picture shows the fit plot of Lrating vs sugars, showing the confidence interval and prediction interval.

The value of sugar in this model is between 0 and 15 grams. Due to the good fit of 59.41%, we have reason to believe that the prediction result is very reliable when performing interpolation prediction.

But if statistical extrapolation is performed (when the value of sugars is greater than 15), then the reliability of the predicted results is questionable.