1. An insurance company wants to estimate the duration of customer service calls. For that purpose, it collects data (records the duration on a sample of customer service calls). Out of this sample the company computes the sample mean (18 minutes) and the sample standard deviation (3 minutes).

a. Assume that the sample size was 20 calls. What would the 95% confidence interval be?
We have a sample size $n = 20$, sample mean $\bar{x} = 18$ (minutes), sample standard deviation $s = 3(minutes)$.

Thus, the 95% confidence interval can be found as $18 \pm 1.96(0.67)$. We are 95% confident that the estimate duration of customer services calls will be between approximately 16.69 and 19.31 minutes, with a margin of error of 0.67 minutes. Hence the 95% confidence range is between 16.69 to 19.31

Reference: https://www.socscistatistics.com/confidenceinterval/default3.aspx

b. Assume now that the company collected much more data (records duration of 5000 calls).
What would the 95% confidence interval be?

Now, if we have a sample size $n = 5000$, sample mean $\bar{x} = 18$ (minutes), sample standard deviation $s = 3(minutes)$.

Thus, the 95% confidence interval can be found as $18 \pm 1.96(0.04)$. We are 95% confident that the estimate duration of customer services calls will be between approximately 16.69 and 19.31 minutes, with a margin of error of 0.04 minutes.

At the same confidence level, a larger sample size will lead to a narrower confidence interval. As the sample size increases, the standard error will decrease, as we have seen from the above example. Hence here the 95% confidence interval range is between 17.92 to 18.08.

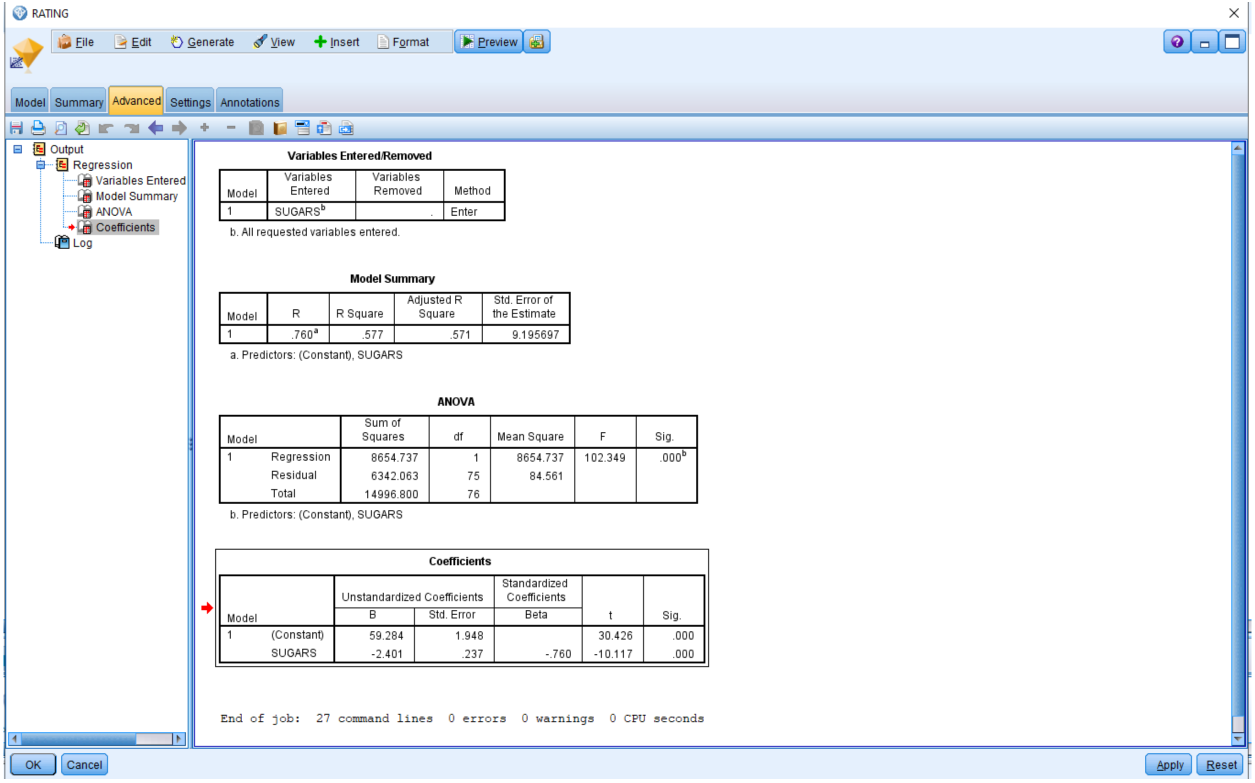
Reference: https://www.socscistatistics.com/confidenceinterval/default3.aspx

Question 2. Use the *cereals* dataset, for the following exercises. Perform a regression analysis using SPSS Modeler to estimate *rating* based on *sugars* alone.

For this part of the question we have given a data set which contains nutritional information and grocery shelf location for 77 breakfast cereals and we have to calculate the following.

a. What is the estimated regression equation? Explain clearly the value of the slope coefficient you obtained in the regression equation.

The estimated regression equation is:
Rating = sugar * - 2.401 + 59.284.



Variables Entered/Removed

Model	Variables Entered	Variables Removed	Method
1	SUGARS[b]	.	Enter

b. All requested variables entered.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.760[a]	.577	.571	9.195697

a. Predictors: (Constant), SUGARS

ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	8654.737	1	8654.737	102.349	.000[b]
	Residual	6342.063	75	84.561		
	Total	14996.800	76			

b. Predictors: (Constant), SUGARS

Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	59.284	1.948		30.426	.000
	SUGARS	-2.401	.237	-.760	-10.117	.000

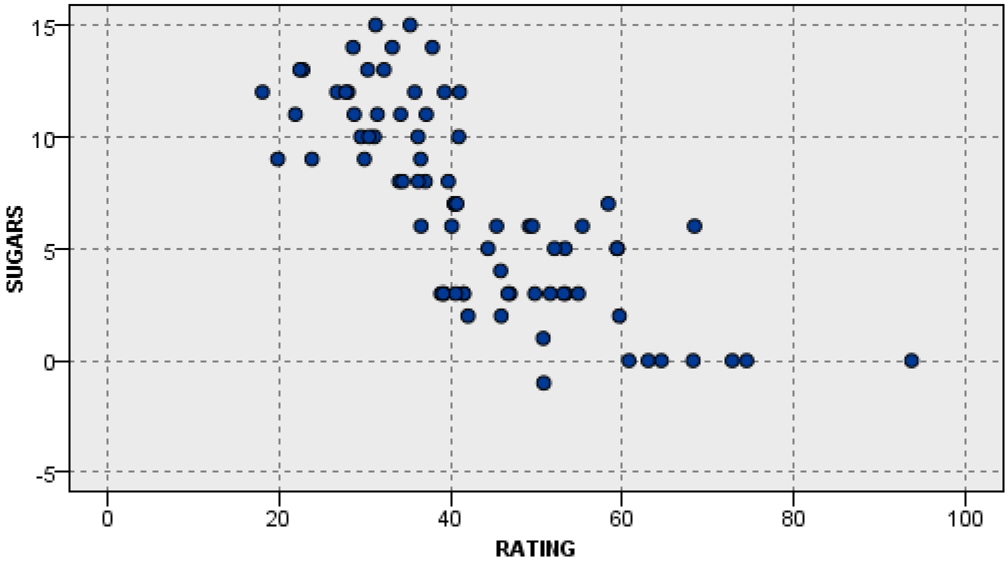
```
End of job:  27 command lines  0 errors  0 warnings  0 CPU seconds
```

b. What does the value of the y-intercept mean for the regression equation you obtained? Does it make sense in this example?

The slope coefficient represents the estimated change in rating per increase in sugar intake in each breakfast cereal from the total of 77 cereals options has a reciprocal or inverse relationship to the rating. As the coefficient of slope is negative 2.401. Which is not surprising as the increase intake of sugar can leads to many deceases and have harmful effect on health in long term.

The y-intercepts represent the estimated rating are zero, which makes sense because it is not unreasonable to say that decrease in sugar consumption can lead to cereal which is the healthiest food and hence have a higher rating.

From the above regression we can also say that the both of the regression coefficients are statically significant. We know that because the significant value obtain from the above regression is 0.000 which indicates that $p < 0.001$. Below is the scatter plot which shows the same with increase in sugar, it leads to lower rating or lower consumption of sugar breakfast has higher ratings.



Scatter plot for the rating and sugar.

c. What would be a typical prediction error obtained from using this model to predict *rating*? Which statistic are you using to measure this?

The predictive power of the model can be explained by standard error of the estimate, of our given case we have standard error of the estimate is 9.19569, which says that we have approximately about 9.19 units (grams of sugar) of typical error we may in counter when we made this prediction. For this assignment we have calculate the prediction error on trained data not over the testing or separated data. We will be doing the prediction of error on the sub set of the data for training and testing purposes in our further assignments.

d. How closely does our model fit the data? Which statistic are you using to measure this?

The R2 values from the above models are 0.577 and the adjusted R square value is also the same 0.577. The above model as a predictor fits the data the best, as it indicates that about 57% of the variations is explained by the regression, as R square is the measure of fit.

e. Find a point estimate for the rating for a cereal with a sugar content of 7 grams.

Estimating the rating for cereals with sugar content of 7 grams can be calculate as below:

Rating = 59.284 - 2.401 (7) = 42.577.

Hence, 42.577 is the rating for the cereals with sugar content of 7 grams.