Assignment 3

Team Based Assignment (100 pts)

The Advertising data set collects sales data of a product in a sample of 200 different markets, along with advertising budgets for the product in each of those markets for three different media: TV, radio, and newspaper. The Advertising data set lists sales, in thousands of units, and TV, radio, and newspaper budgets, in thousands of dollars.

Question 1. (30 pts)

Let's imagine that you want to estimate the average number of units sold for the full population (all markets).

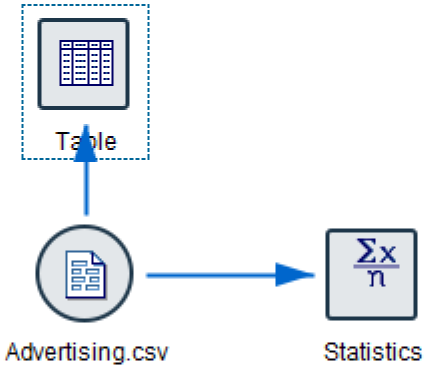
i. What would the 95% confidence interval be?

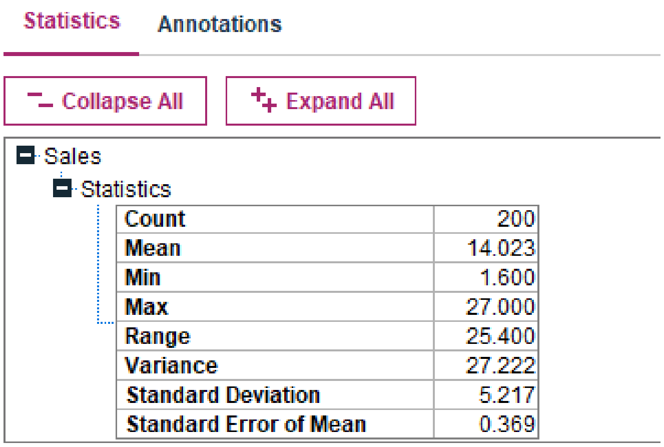
The confidence interval for a population is defined as sum of point estimate and margin of error. 95% of the confidence interval implies that there is 95% probability of the estimated parameter to be contained within this calculated limit.

The confidence interval for estimating average number of units sold for full population is calculated by the formula μ=X−±z.SE (as the sample size is $\geq$ 30)

Where X is sample mean, z is z score for 95% of confidence interval which is ±1.96 and SE is standard error of mean (=sample mean/sqrt(sample size)).

From the below figure, after executing the statistics node in SPSS modeler, the above statistics are calculated.





Sales

Statistics

Count	200
Mean	14.023
Min	1.600
Max	27.000
Range	25.400
Variance	27.222
Standard Deviation	5.217
Standard Error of Mean	0.369

From the formula estimating average number of units sold for full population,

=>μ= 14.023±1.96*0.369

95% lower confidence limit = 13.299and

95% upper confidence limit = 14.746

This means that with a confidence level of 95% the population average sales units in all markets lie between 13.299 units and 14.746 units

ii. What would happen to the confidence interval at the same confidence level if you had a larger sample? Explain it qualitatively (you don't have to provide a numerical value as answer)

Note: you can do this manually, using the tables in lecture notes 03a (you don't need Modeler).

As the sample size increases, at the same confidence level, the width of the confidence interval decreases. Theoretically, the big picture is easier to understand when there is clear data with a greater number of samples. Hence, the probability of estimating the population parameters increases when there are more samples. Practically, the standard deviation or the standard error of mean is inversely proportional to the square root of sample size. So in the formula μ=X−±z.SE where SE = σ / n or SE = s / n, with the effect of ± in estimating μ, the width decreases.

Also, the same is observed in t table. In the t table, at the same confidence level, the t values decrease with the increase in the sample size, which effects a decrease in the width for estimating the population parameters. Thus, there is a decrease in the margin of error with a decrease in confidence level (which is not recommended as the interval decreases) and an increase in sample size, which is mostly recommended.

Question 2. (70 pts)

Perform three regression analyses using SPSS Modeler to estimate Sales based on:

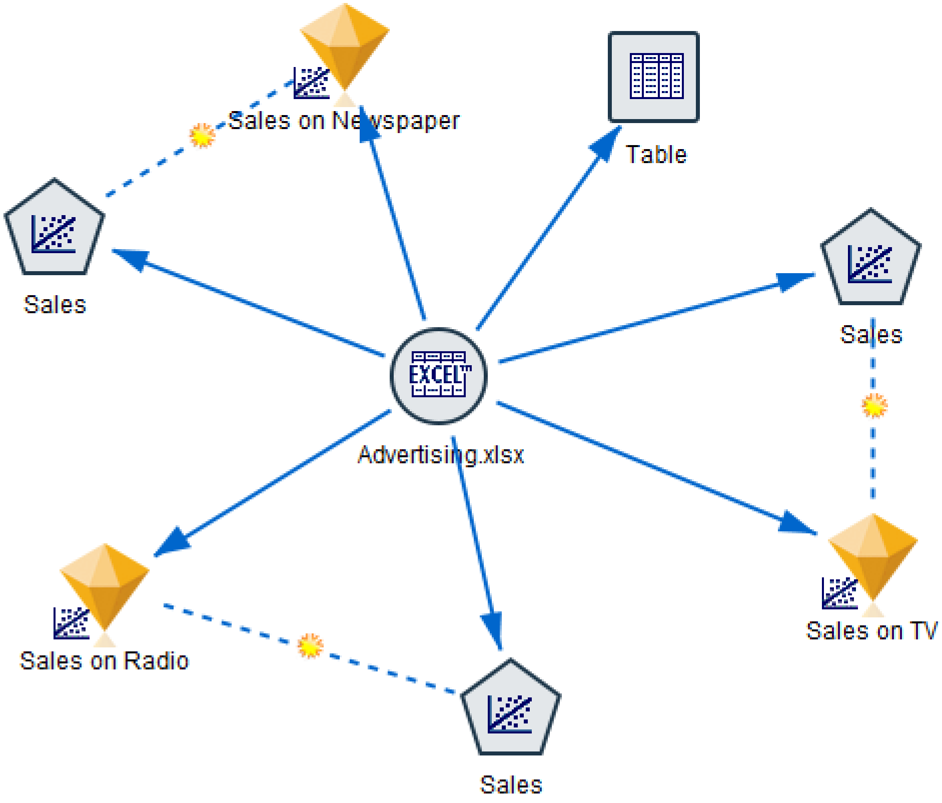
a) TV alone;

b) Newspaper alone;

c) Radio alone

i. What is the estimated regression equation for each regression analysis (a, b and c)?

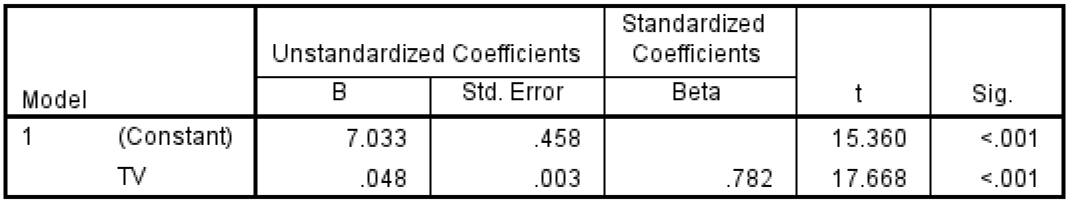
After running the following regression model in SPSS modeler, the output is as shown in below tables



The regression equation for estimating sales based on advertising budget of TV is given from the coefficients table; with an intercept of 7.033 and a slope of 0.048

=> sales equation from TV = 7.033+0.048*TV

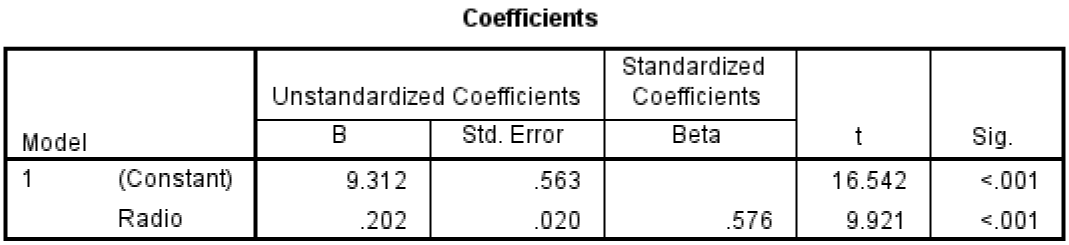
Coefficients



Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	7.033	.458		15.360	<.001
	TV	.048	.003	.782	17.668	<.001

The regression equation for estimating sales based on advertising budget of Radio is given from the coefficients table; with an intercept of 9.312 and a slope of 0.202

=> sales equation from Radio = 9.312+0.202*Radio

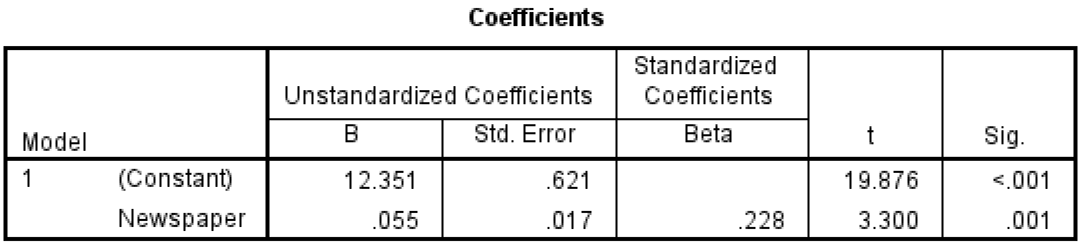


Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	9.312	.563		16.542	<.001
	Radio	.202	.020	.576	9.921	<.001

Similarly, the regression equation for estimating sales based on advertising budget of Newspaper is given from the coefficients table; with an intercept of 12.351 and a slope of 0.55

=> sales equation from Newspaper = 12.351+0.55*Newspaper



Coefficients

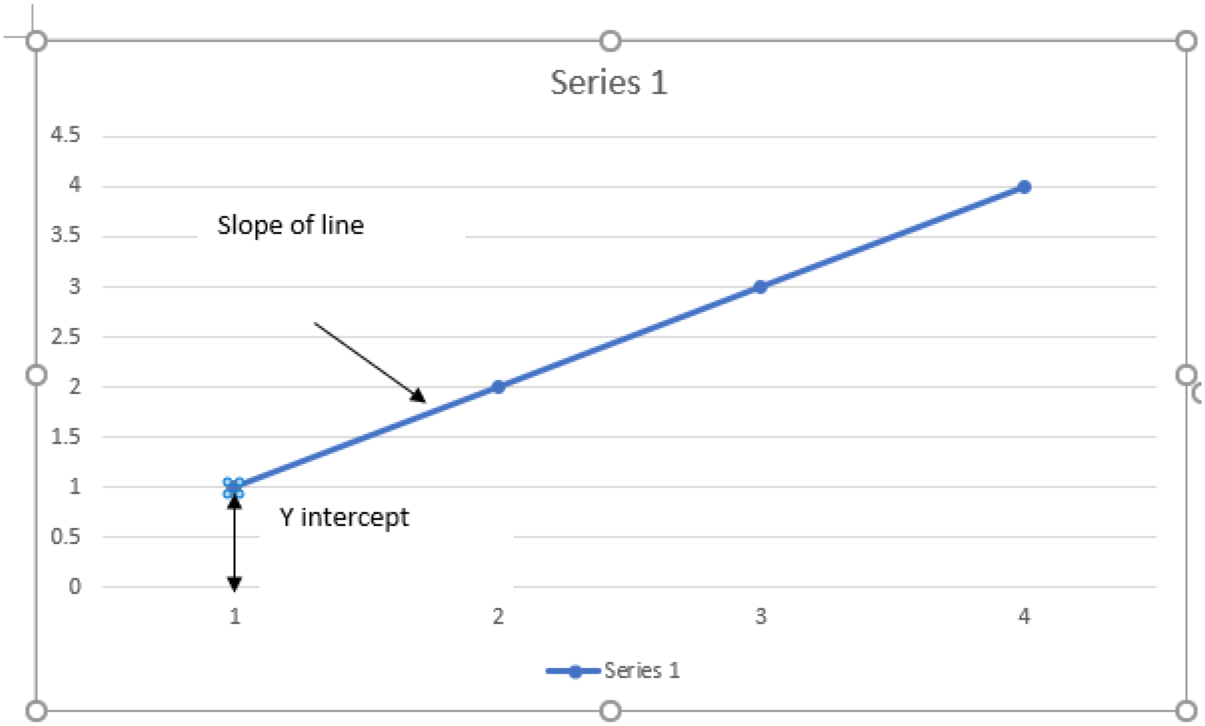
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	12.351	.621		19.876	<.001
	Newspaper	.055	.017	.228	3.300	.001

ii. Explain clearly the value of the slope coefficient you obtained in each of the three regression equations. What does the value of the y-intercept mean for the three regression equations you obtained? Does it make sense in each of the three cases?

The y-intercept and the slope coefficient in the regression equation do make sense and have specific interpretations. The regression equation is generally of the form $Yi- = b0 + b1Xi1$ where b_1 is the slope and b_0 denotes the y-intercept. The significance of b_1 it is the estimate change in Y by b_1 for each unit increase in X_1. That is, the slope gives the variation in y intercept when there is a unit increase in the mean of sample set X_{i1}. And the y intercept is average value of y when X_1 is zero. That is, it gives the mean of observation of the population when sample population mean is zero.

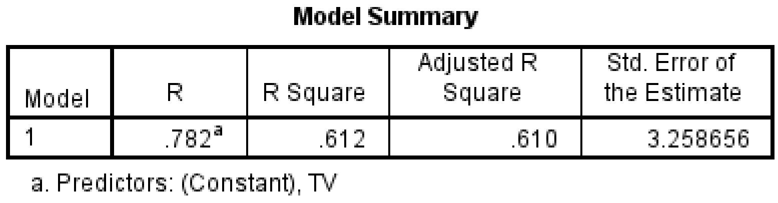
Whether the regression equation in either in above form or μ=X−±z.SE, the denotations change but the meaning of slope and intercept doesn't. The slope and y intercept denote the characteristics of the relationship between the variables sample mean and population mean in the above two linear equations.

In the regression equations of above question, sales equation from TV = 7.033+0.048*TV, sales equation from Radio = 9.312+0.202*Radio and sales equation from Newspaper = 12.351+0.55*Newspaper the slope coefficients 0.048, 0.202 and 0.55 denotes the slope of the linear equation and the intercepts 7.033, 9.312 and 12.351 gives the distance of point y from the origin.



iii. What would be a typical prediction error obtained from using the regression model to predict sales when using TV alone as a predictor? Which statistic are you using to measure this?

Standard error of estimate gives the prediction error which may be encountered when making predictions while using a regression equation. From the below model summary tables while predicting the sales with an input of TV, Radio and newspaper, standard error of the estimate is 3.26, 4.27 and 5.09 respectively, which signifies that the prediction can make an error of around 3.26, 4.27 and 5.09 units for respective regression equations.



Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.782[a]	.612	.610	3.258656

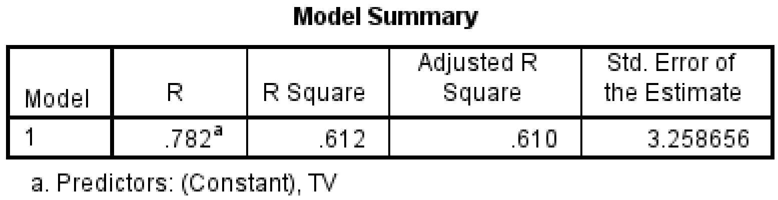
a. Predictors: (Constant), TV

While using Tv alone as a predictor, a typical prediction error of 3.26 units in sales can be observed around the predicted sales data.

iv. How closely do the three models fit the data? Which statistic are you using to measure this?

Based on the R square parameter, the goodness of fit is determined. R square shows how well the prediction sits on the actual data value. A high R square model fits the model better than low R square, which means that 0 gives the least fit and 1 being better fit with the actual data.

In case of sales from Tv budgeting, a fit of 61.2% is expected. It tells that the 61.2% of the variation of dependent target sales around its mean.

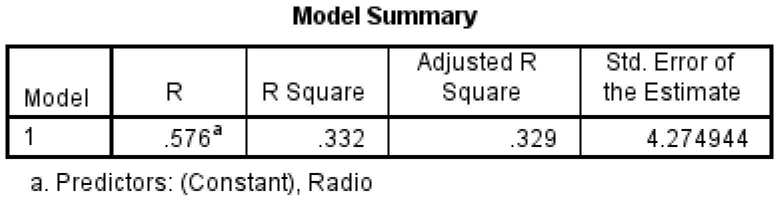


Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.782[a]	.612	.610	3.258656

a. Predictors: (Constant), TV

In case of sales from Radio budgeting, a fit of 33.2% is expected. It tells that the predicted value from the regression coefficients fits the actual value with a precision of 33.2%, shown as R square in below table

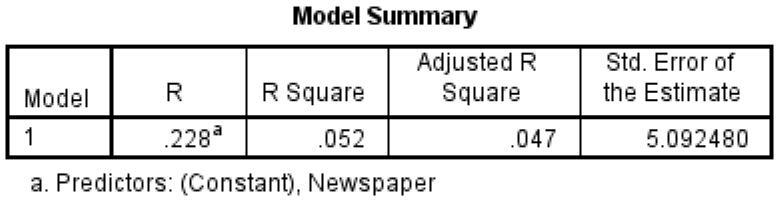


Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.576[a]	.332	.329	4.274944

a. Predictors: (Constant), Radio

In case of sales from Tv budgeting, a fit of 5.2% is expected. It tells that the predicted value from the regression coefficients fits the actual value with a precision of 5.2%, shown as R square in below table



Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.228[a]	.052	.047	5.092480

a. Predictors: (Constant), Newspaper

Depending on the above three R squares, the order of best fit goes from sales on Tv, Radio and least fit is Newspaper model. Hence the acceptance of regression model for newspaper should be given least priority than Radio and TV models.

v. Find a point estimate for the Sales with a TV advertising expenditure of 100 ($100,000).

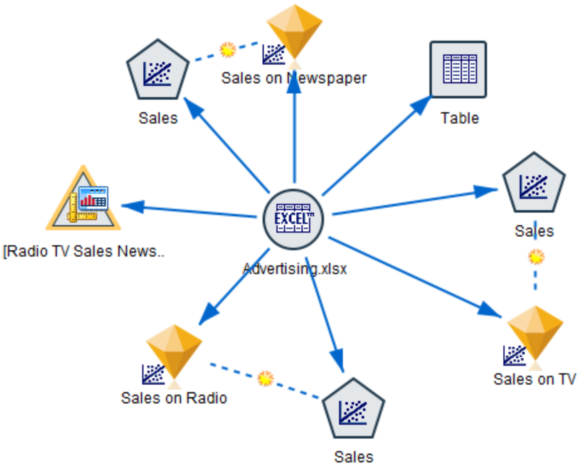
From the regression equation for estimating the sales based on budgeting of TV advertising, we get, sales equation from TV = 7.033+0.048*TV

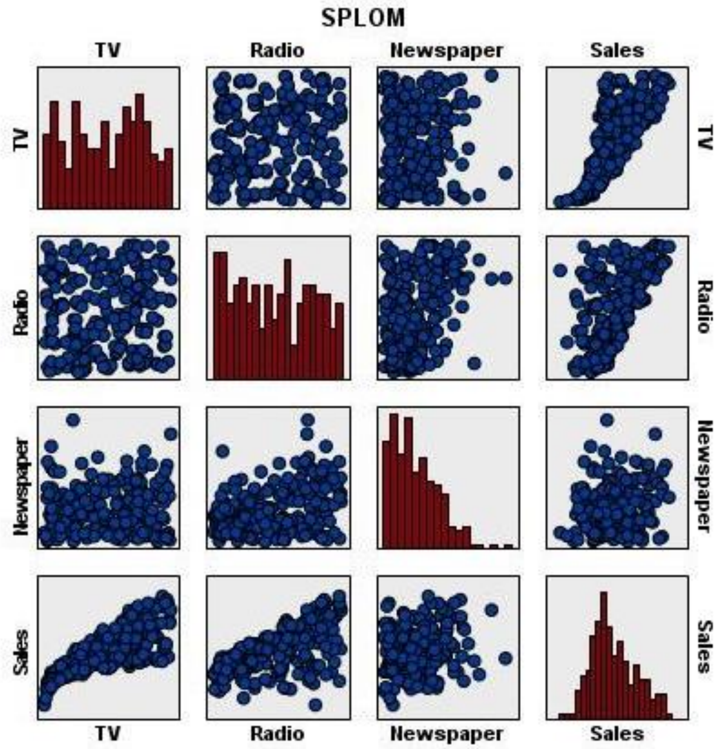
=> point estimate of sales with TV advertising expenditure of 100 is given by 7.033+0.048*100 = 11.833 in thousands of units. Also, there could possibly be a random error of ‘e’ as we do not have full population data.

From the above standard error of estimate, a unit of 3.26 can be considered as an error in total predicted value from 11.833 units which could possibly be 8.573 and 15.093 units

vi. Plot a scatterplot matrix chart to graphically depict the correlation between pairs of variables

When a graph plot is attached to the sales data file and run by selecting all the variables, the following scatterplot matrix chart is an output. It gives the correlation between each set of variables with the other set of variables. For example, one can read from the following figure that the data of Tv and radio is more or less equally distributed whereas newspaper is mostly negatively skewed, and the sales data is normally distributed. The radio and TV data is scattered all through while newspaper and TV data is mostly negatively skewed and there is more or less a linear relationship between sales and TV.


Sales on Newspaper
Sales
Table
[Radio TV Sales News..
EXCEL
Advertising.xlsx
Sales
Sales on TV
Sales on Radio
Sales


SPLOM
TV
Radio
Newspaper
Sales
TV
Radio
Newspaper
Sales
TV
Radio
Newspaper
Sales
TV
Radio
Newspaper
Sales