North South University, School of Business
MKT 631 Marketing Research
Instructor: Mahmood Hussain, PhD

Data Analysis for Marketing Research - Using SPSS

Introduction

In this part of the class, we will learn various data analysis techniques that can be used in marketing research. The emphasis in class is on how to use a statistical software(SAS, SPSS, Minitab, SYSTAT, and so on) to analyze the data and how to interpret the results in computer output. The main body of the data analysis part will be devoted to multivariate analysis when we try to analyze two or more variables.

Classification

In multivariate data analysis(i.e., when we have two or more variables) we are trying to **find** the **relationship** between the variables and **understand** the **nature of the relationship**. The choice of the techniques primarily depends on **what measurement scales** we used for the variables and in **which way** we are trying to find the relationship between the variables. For example, we can see the relationship between variable x and y, by examining the **difference** of y at each value of x or by examining the **association** between the two variables. In class, basically the two classes of data analysis techniques will be taught: analysis of difference and analysis of association.

I. Analysis of Difference

Ex. t-test(z-test), Analysis of Variance(ANOVA)

II. Analysis of Association

Ex. Cross-tabulation, Correlation, Regression

Suggestion

Before you start any of the data analysis techniques, check:

1. What are the variables in my analysis?
2. Can the variables be classified as dependent and independent variable?
3. If yes, what is (ind)dependent variable?
4. What measurement scales were used for the variables?

Data Analysis I: Cross-tabulation (& Frequency)

In cross-tabulation, there's no specific dependent or independent variables. The cross-tabulation just analyzes whether the two(or more) variables measured with nominal scales are associated or equivalently, whether the cell counts of the values of one variable are different across the different values(categories) of the other variable. All of the variables are measured by nominal scale.

Example 1

Research objective: To see whether the preferred brands(brand A, brand B, and brand C) are associated with the locations(Denver and Salt Lake City); Is there any difference in brand preference between the two locations?

Data: We selected 40 people in each city and measured what their preferred brands were.
If the preferred brand is A, the 'favorite brand' =1.
If the preferred brand is B, the 'favorite brand' =2.
If the preferred brand is C, the 'favorite brand' =3.

Denver

ID	favorite brand	ID	favorite brand	ID	favorite brand	ID	favorite brand
1	1	11	1	21	1	31	3
2	3	12	2	22	3	32	3
3	3	13	2	23	3	33	2
4	1	14	2	24	3	34	3
5	3	15	3	25	1	35	1
6	2	16	3	26	3	36	3
7	3	17	1	27	3	37	3
8	3	18	3	28	3	38	3
9	3	19	1	29	3	39	2
10	1	20	2	30	3	40	1

Salt Lake City

ID	favorite brand	ID	favorite brand	ID	favorite brand	ID	favorite brand
1	2	11	1	21	3	31	2
2	1	12	1	22	3	32	2
3	2	13	1	23	2	33	1
4	1	14	1	24	3	34	2
5	2	15	2	25	2	35	1
6	1	16	2	26	1	36	1
7	2	17	1	27	2	37	1
8	1	18	2	28	3	38	1
9	2	19	1	29	3	39	1
10	1	20	2	30	2	40	1

Procedures in SPSS

1. Open SPSS. See upper-left corner. You see 'SPSS Data Editor'.

2. Type in new the data(Try to save it!). Each column represents each variable.
   Or, if you already have the data file, just open it.(***data_crosstab_1.sav***)

3. Go to top menu bar and click on "**Analyze".** Then, now you've found **"Descriptive Statistics".** Move your cursor to "Descriptive Statistics"

4. Right next to "Descriptive Statistics", click on "**Crosstabs…**". Then you will see:

These are the variable names and can be changed for different dataset.

5. Move your cursor to the first variable in upper-left corner box and click. In this example, it's 'Location'. Click to the ▶ button just left to 'Row(s):' box. This will move 'Location' to the 'Row(s):' box. In the same way, move the other variable 'Brand' to 'Column(s):' box. You have:

6. Go to [Statistics...] in the bottom and click. You will see another menu screen is up. In that menu box, click on "Chi-square" and "Contingency coefficient" as follows:

and click on [Continue] on right-upper corner.

7. You've been forwarded to the original menu box. Click on [OK] in right-upper corner.

8. Now you've got the final results at **Output- SPSS Viewer** window as follows:

Crosstabs

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
LOCATION * BRAND	80	100.0%	0	.0%	80	100.0%

BRAND * LOCATION Crosstabulation

Count

		LOCATION		Total
		1	2	
BRAND	1	10	19	29
	2	7	16	23
	3	23	5	28
Total		40	40	80

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	17.886[a]	2	.000
Likelihood Ratio	18.997	2	.000
N of Valid Cases	80		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 11.50.

Symmetric Measures

		Value	Approx. Sig.
Nominal by Nominal	Contingency Coefficient	.427	.000
N of Valid Cases		80	

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

<u>Interpretation</u>

First table simply shows that there are 40 observations in the data set. Recall the dataset again. There were 20 observations in each city.

From the second table, it is found that among 40 people in Denver(location=1), 10, 7, and 23 people prefer the brand A, B, and C, respectively. On the other hand, 19, 16, and 5 out of 40 people in Salt Lake City (location=2) like brand A, B, and C, respectively. From the third table, the chi-square value is 17.89(`Chi-Square = 17.886`) and the associated p-value for this chi-square value is is 0.00(`Asymp. Sig. = 0.000),` which is less than 0.05. Therefore, we conclude that people in different city prefer the different brand or that consumers' favorite brand is associated with where they live in. (This conclusion is also supported by the final table that shows the contingency coefficient is 0.427.)

Looking at the column totals or row totals in the second table, we can **also** get the **frequency** for each variable. In this example, among all 80 people in my sample, 29, 23, and 28 people said that their most favorite brands are A, B, and C, respectively. And 40 people are selected from Denver (location=1) and the other 40 are from Salt Lake City (location=2).

Example 2

49ers vs Packers

Research objective: To see whether winning (or losing) the basketball game is associated with whether the game is home-game or away-game.

Data: We selected the data of the basketball game between 49ers and Packers for the period between 19XX and 19&&

Game at (1=SF, 2=GB)	Game results (1=49ers win, 2=49ers lose)
1	1
2	1
1	1
2	2
1	1
2	2
1	1
2	2
1	1
2	1
1	2
2	2
1	1
2	2
1	1
2	2
1	1
2	1
1	2
2	1
1	1
2	2
1	1
2	2
1	1
2	2
1	2
2	2
1	1
2	2

Results

GAME_AT * RESULT Crosstabulation

Count

		RESULT		Total
		1	2	
GAME_AT	1	12	3	15
	2	4	11	15
Total		16	14	30

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	8.571[b]	1	.003		
Continuity Correction[a]	6.563	1	.010		
Likelihood Ratio	9.046	1	.003		
Fisher's Exact Test				.009	.005
Linear-by-Linear Association	8.286	1	.004		
N of Valid Cases	30				

a. Computed only for a 2x2 table

b. 0 cells (.0%) have expected count less than 5. The minimum expected count is 7.00.

Symmetric Measures

		Value	Approx. Sig.
Nominal by Nominal	Contingency Coefficient	.471	.003
N of Valid Cases		30	

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

Interpretation

Among total 30 games, 49ers won 16 games and lost 14 games. Each teams tends to win the game when the football game is "home-game" for each team:

Among 16 49ers' victories, 12 games were played at SF.
Among 14 Packers' victories, 11 games were played at GB.

The chi-square for this table is 8.571(`Pearson Chi-Square = 8.571`) and the p-value is 0.003 (`Asymp. Sig. = 0.003`). Since p-value = 0.003 is less than 0.05(we typically assume that our pre-specified level of significance is 0.05) and contingency coefficient is far from '0'(.471), we conclude that the likelihood that a school wins the football game is associated with whether the game is 'home-game' or 'away-game' for the school.

Data Analysis II: t-test(or z-test)

Dependent variable: metric scale(interval or ratio scale)
Independent variable: nonmetric scale(nominal)

T-test(or Z-test) tests whether there's a significant difference in dependent variable between **two groups**(categorized by independent variable).

Example

Research objective: To see whether the average sales of brand A are different between when they offer a price promotion and when they do not.

Data: We measured sales of brand A when we offer promotion(promotion =1) and regular price(promotion = 2).

Promotion	Sales	Promotion	Sales
1	75	2	51
1	87	2	70
1	83	2	37
1	45	2	62
1	95	2	90
1	89	2	72
1	74	2	45
1	110	2	78
1	75	2	45
1	84	2	76

Dependent variable: Sales
Grouping variable: Promotion(1 = Yes, 2 = No)

Procedures in SPSS

1. Open SPSS. See upper-left corner. You see in 'SPSS Data Editor'.

2. Type in new the data(Try to save it!). Each column represents each variable.
   Or, if you already have the data file, just open it.(***data_t-test_1.sav***)

3. Go to top menu bar and click on "**Analyze".** Then, now you've found **"Compare Means".** Move your cursor to "Compare Means"

4. Right next to "Compare Means", click on "**Independent Samples T-Test…**". Then you will see:

These are the variable names and can be changed for different dataset.

5. Move your cursor to the dependent variable in left-hand side box and click. In this example, it's 'Sales'. Click to the ▶ button just left to 'Test Variable(s):' box. This will move 'Sales' to the 'Test Variable(s):' box. Then move the independent variable(grouping variable) to 'Grouping variable:'. In this example, it's 'Promo'. Click on 'Define Groups', and specify what numbers are for what group in your data by typing the numbers into the boxes.

Then, click on Continue. You have:

6. Click on OK in right-upper corner.

7. Now you've got the final results at **Output- SPSS Viewer** window as follows:

Results

T-Test

Group Statistics

	PROMO	N	Mean	Std. Deviation	Std. Error Mean
SALES	1	10	81.70	16.872	5.336
	2	10	62.60	17.386	5.498

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
SALES	Equal variances assumed	.458	.507	2.493	18	.023	19.10	7.661	3.004	35.196
	Equal variances not assumed			2.493	17.984	.023	19.10	7.661	3.003	35.197

Interpretation

From the first table, the mean sales under the promotion and non-promotion are 81.7 and 62.6, respectively. The test statistic, t, for this observed difference is 2.49(`t= 2.493`). The p-value for this t-statistic is 0.023(`Sig.(2-tailed)=0.023`). Since p-value (0.023) is less than 0.05, we reject the null hypothesis and conclude that there's a significant difference in average sales between when firms offer price promotion and when they offer just regular prices.

Data Analysis II: Analysis of Variance (One-way ANOVA)

Dependent variable: metric scale(interval or ratio scale)
Independent variable: nonmetric scale(nominal)

ANOVA(F-test) tests whether there's a significant difference in dependent variable among **multiple groups**(categorized by independent variable).

Example

Research objective: There are three training programs of sales person. We want to see whether the different job training methods for the employees result in different level of job satisfaction.

Data: We trained 15 rookies in the company with different program. Each training method is applied to 5 people.

ID #	Trained by [a]	Job satisfaction level [b]
1	1	5
2	1	2
3	2	3
4	2	4
5	3	5
6	1	6
7	2	2
8	3	5
9	2	2
10	2	3
11	1	5
12	1	7
13	3	4
14	3	4
15	3	5

[a] Grouping variable: training program(1= trained by program #1, 2= trained by program #2, 3 = trained by program #3)
[b] Dependent variable: Job satisfaction (1: Strongly dissatisfied --- 7: strongly satisfied)

Procedures in SPSS

1. Open SPSS. See upper-left corner. You see in 'SPSS Data Editor'.

2. Type in new the data(Try to save it!). Each column represents each variable.
   Or, if you already have the data file, just open it.(***data_anova_1.sav***)

3. Go to top menu bar and click on "**Analyze".** Then, now you've found **"Compare Means".** Move your cursor to "Compare Means".

4. Right next to "Compare Means", click on "**One-way ANOVA…**". Then you will see:

These are the variable names and can be changed for different dataset.

5. Move your cursor to the dependent variable in left-hand side box and click. In this example, it's 'Satis'. Click to the ▶ button just left to 'Dependent List:' box. This will move 'Satis' to the 'Dependent List:' box. Then move the independent variable(grouping variable) to 'Factor:'. In this example, it's 'Program'.

6. Click on [Options]. A new sub-menu box will appear. Then, click on the following boxes:

and click on [Continue] in right-upper corner.

7. Click on [OK] in right-upper corner.

8. Now you've got the final results at **Output- SPSS Viewer** window as follows:

Results

Descriptives

SATIS

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
1	5	5.00	1.871	.837	2.68	7.32	2	7
2	5	2.80	.837	.374	1.76	3.84	2	4
3	5	4.60	.548	.245	3.92	5.28	4	5
Total	15	4.13	1.506	.389	3.30	4.97	2	7

ANOVA

SATIS

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	13.733	2	6.867	4.578	.033
Within Groups	18.000	12	1.500		
Total	31.733	14			

Interpretation

(1) From the first table, the mean satisfaction of the subjects in group 1(i.e., those who were trained by program 1), group 2, and group 3 are 5.0, 2.8, and 4.6, respectively.

(2) Checking ANOVA table, the second output:

Source	SS	DF	MS	F	Sig.
Between	SS_G	df_G	MS_G	F	*p-value*
Within	SS_E	df_E	MS_E		
Total	df_T	SS_T			

$SS_T = SS_G + SS_E$ 31.73 = 13.73 + 18.00

df_T = number of respondents – 1 = 15 – 1 = 14

df_G = number of groups – 1 = 3 – 1 = 2

$df_E = df_T - df_G$ = 14 – 2 = 12

$MS_G = SS_G/df_G$ 6.87 = 13.73 / 2

$MS_E = SS_E/df_E$ 1.50 = 18.00 / 12

$F = MS_G/MS_E$ (*Test statistic*) 4.58 = 6.87 / 1.50

→ As F increases, we are more likely to reject the null hypothesis(H_o).

(3) Conclusion through hypotheses testing

In the second table, since p-value (0.033) is less than 0.05, we reject the null hypothesis and conclude that **there's a significant difference in job satisfaction level among the groups of sales persons who were trained by different program**.

Data Analysis III: Correlation Analysis

1. Variables variable 1 : metric scale
   variable 2 : metric scale

2. What is correlation?

   Correlation measures :
   (1) whether two variables are linearly related to each other,
   (2) how strongly they are related, and
   (3) whether their relationship is negative or positive.

   These can be answered by getting **correlation coefficient**.

3. Correlation coefficient

   (1) How to get correlation coefficient?

$$Corr(X,Y) = \frac{Cov(X,Y)}{\sigma_X \cdot \sigma_Y}$$

$Corr(X,Y)$: Correlation coefficient between variable X and variable Y
$Cov(X,Y)$: Covariance between variable X and variable Y
σ_x : standard deviation of variable x
σ_y : standard deviation of variable Y

cf. Sample Correlation coefficient (r)

$$r = \frac{S_{XY}}{S_X \cdot S_Y} \quad \text{or} \quad r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \cdot \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

S_{XY}: sample covariance between variable X and variable Y

S_x : sample standard deviation of variable X = $\sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}}$

S_y : sample standard deviation of variable Y = $\sqrt{\frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1}}$

n : sample size

(2) Properties of Corr(X,Y)

- $-1 \leq \text{Corr}(X,Y) \leq 1$

- Corr(X,Y) < 0 ➔ X and Y are negatively correlated
  Corr(X,Y) = 0 ➔ X and Y are **UN**correlated
  Corr(X,Y) > 0 ➔ X and Y are positively correlated

- Corr(X,X) or Corr(Y,Y) = 1

(3) Graphical illustration of Corr(X,Y)

Corr(X,Y) > 0

Corr(X,Y) < 0

Example

Research objective : Study of the relationship between age and soft drink consumption

Data: We selected 10 people and obtained their age (variable X) and average weekly soft drink consumption (variable Y).

Respondent #	X	Y*
1	40	84
2	20	120
3	60	24
4	50	36
5	45	72
6	30	84
7	25	96
8	35	48
9	30	60
10	45	36

*** oz.**

Exercise

(1) Using the formula on page 17, get the sample correlation coefficient(r) by hand. (To calculate 'r', you will have to calculate the sample mean of X and the sample mean of Y, first.)

(2) Run the correlation analysis in SPSS.

a) Plot the data (X and Y). Is the plot consistent with the sample correlation coefficient you calculated in (1)?

b) Compare the sample correlation coefficient in computer outputs with (1) and check if the two coefficients are equal.

<u>Procedures in SPSS</u>

1. Open SPSS. See upper-left corner. You see in 'SPSS Data Editor'.

2. Type in new the data(Try to save it!). Each column represents each variable.
   Or, if you already have the data file, just open it.(***data_corr_1.sav***)

[To get the correlation coefficient]

3. Go to top menu bar and click on "**Analyze".** Then, now you've found **"Correlate".** Move your cursor to "Correlate"

4. Right next to "Correlate", click on "**Bivariate..**". Then you will see:

These are the variable names and can be changed for different dataset.

5. Move your cursor to the variable box on your left and click the variables you want to get the correlation coefficient. In this example, they are 'x' and 'y'. Click the ▶ button and move the variables to the right-hand side box

6. Click on the box of ☑ Pearson.

7. Click on OK in right-upper corner.

8. Now you've got the final results at **Output- SPSS Viewer** window as follows:

Correlations

Correlations

		X	Y
X	Pearson Correlation	1	-.833**
	Sig. (2-tailed)	.	.003
	N	10	10
Y	Pearson Correlation	-.833**	1
	Sig. (2-tailed)	.003	.
	N	10	10

**. Correlation is significant at the 0.01 level

Interpretation

Consumer's weekly soft drink conumption and their age are negatively correlated (`-0.833`). Therefore, we conclude that younger consumers tend to consume soft drink more.

[To get the plot of (X, Y)]

4. Go to top menu bar and click on "**Graphs".** Then, now you've found **"Scatter..".**

5. Click on "Scatter.." Then, you will be introduced to a new sub-menu box.

6. Choose "Simple" and click Define .

7. Click on OK in right-upper corner.

8. Now you've got the final results at **Output- SPSS Viewer** window as follows:

Results

Interpretation

The pattern in the plot shows that, as X increases, Y decreases. This indicates a ‘negative’ relationship between X and Y. Recall that the correlation coefficient we got from the other output was –0.833, which is consistent with this plot.

Data Analysis III: Regression Analysis

[1] Simple Regression

1. Variables: variable 1: Dependent(Response) variable: metric scale
   variable 2: Independent(Predictor) variable: metric or nominal scale

2. Purpose

   (1) understanding / explanation
   (2) prediction

   ↓

   How can we achieve these two goals?

   ↓

   Getting $\boxed{Y = \beta_o + \beta_1 X_1}$

   ↓

(1) estimate coefficient(s)-β_o and β_1 (= estimate regression function)
(2) show whether the coefficients that we estimated are significant [hypothesis testing]
(3) know how useful is the regression function ? [goodness of fit]

3. What to do in regression analysis

(1)-1 How to estimate parameters(β_o and β_1): = How to estimate regression function
Least Square method
The estimated values are given from computer output.

(2) Are the coefficients significant? Answer: p-value

(3) How strong is the relationship between variables? Ans: R^2
Is such relationship really meaningful(significant)? Ans: F-test

Example

Research objective: To see whether advertising frequency affects the sales

Data

Y	X1
97	45
95	47
94	40
92	36
90	35
85	37
83	32
76	30
73	25
71	27

Y = sales X1 = advertising frequency

Procedures in SPSS

1. Open SPSS. See upper-left corner. You see in 'SPSS Data Editor'.

2. Type in new the data(Try to save it!). Each column represents each variable.
   Or, if you already have the data file, just open it.(***data_reg_1.sav***)

3. Go to top menu bar and click on "**Analyze".** Then, now you've found **"Regression".** Move your cursor to "Regression"

4. Right next to "Regression", click on "**Linear..**". Then you will see:

5. Move your cursor to the variable box on your left and click the dependent variable(in this example, it is ‘y’.) and move it to ‘Dependent:’ box by clicking the ▶ button. Similarly, move independent variable(‘x1’) to ‘Independent(s):’ box by clicking the ▶ button.

6. Choose Method: Enter . You’ve got:

7. Click OK in right-upper corner.

8. Now you’ve got the final results at **Output- SPSS Viewer** window as follows:

<u>Results</u>

Regression

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	X1[a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: Y

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.922[a]	.850	.831	3.927

a. Predictors: (Constant), X1

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	697.004	1	697.004	45.188	.000[a]
	Residual	123.396	8	15.424		
	Total	820.400	9			

a. Predictors: (Constant), X1

b. Dependent Variable: Y

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	42.509	6.529		6.510	.000
	X1	1.217	.181	.922	6.722	.000

a. Dependent Variable: Y

<u>Interpretation</u>

(1) From the last table, estiamted regression coefficients are : β_o (Constant)= 42.5 and β_1 (coefficient for X1)= 1.22. The p-value for testing Ho: $\beta_1 = 0$ is 0.000 (see the last column: Sig. = .000). Therefore, we reject the null hypothesis and we conclude that sales(Y) is significantly affected by advertising frequency(X1). Based on the estimated regression equaition Y=42.5+1.22*X1, we expect that the sales(Y) will increase by 1.22 units if we increase advertising frequency(X1) by one unit.

(2) From the second table, $R^2 = 0.85$ (or 85%)(see R Square = .850). This indicates that the 85% of total variance of sale(Y) is explained by the estimated regression equation or, generally, by the advertising frequency(X1). Look at the third table, ANOVA table.

Source	SS	DF	MS	F	Sig
Regression	SS_R	df_R	MS_R	F	*p-value*
Residual	SS_E	df_E	MS_E		
Total	SS_T	df_T			

$SS_T = SS_R + SS_E$ ➔ 823.4 = 697.0 + 123.4

df_T = number of respondents – 1 = 10 – 1 = 9

df_R = number of independent variable = 1

$df_E = df_T - df_R$ = 9 – 1 = 8

$R^2 = SS_R/SS_T$ = 697.00 / 820.40 = 0.85.

$MS_R = SS_R/df_R$ 697.00 = 697 / 1

$MS_E = SS_E/df_E$ 15.42 = 123.40 / 8

$F = MS_R/MS_E$ (*Test statistic*) 45.19 = 697.0 / 15.42

→ As F increases, we are more likely to reject the null hypothesis(H_o).

P-value associated for this F-statistic is 0.000. Therefore, we conclude that the current regression equation meaningfully explains the relationship between the sales(Y) and advertising frequency(X1). (Note: there are two types of p-values in the output-one in regression coefficient table and the other in ANOVA table).

[2] Multiple Regression

1. Variables: variable 1: Dependent(Response) variable: metric scale
   variable 2: Independent(Predictor) **variables**: metric or nominal scale

2. Purpose [the same as in simple regression]

   (1) understanding / explanation
   (2) prediction

   ↓

   How can we achieve these two goals ?

   ↓

   Getting $\boxed{Y = \beta_o + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k}$

   ↓

   (1) estimate coefficient(s) (or, equivalently, estimate the regression function)
   (2) show whether the coefficients that we estimated are significant
   (3) know how useful is the regression function ? [goodness of fit]
   (4) find what variable is relatively more important

3. What to do in multiple regression analysis

(1) How to estimate parameters($\beta_o, \beta_1, \ldots, \beta_k$): = How to estimate the regression function
Least Square method
The estimated values are given at computer output.

(2) Are the coefficients significant? Answer. `p-value`

(3) How strong is the relationship between variables? Answer: R^2
Is such relationship really meaningful(significant)? Answer: F-test

Example(continued)

Research objective: To see whether advertising frequency(X1) and the # of sales persons(X2) affect the sales(Y)

Data

Y	X1	X2
97	45	130
95	47	128
94	40	135
92	36	119
90	35	124
85	37	120
83	32	117
76	30	112
73	25	115
71	27	108

Y = sales, X1 = advertising frequency, X2 = # of sales persons

Procedures in SPSS

1. Open SPSS. See upper-left corner. You see in 'SPSS Data Editor'.

2. Type in new the data(Try to save it!). Each column represents each variable.
   Or, if you already have the data file, just open it.(***data_reg_2.sav***)

3. Go to top menu bar and click on "**Analyze".** Then, now you've found **"Regression".**
   Move your cursor to "Regression"

4. Right next to "Regression", click on "**Linear..**". Then you will see:

5. Move your cursor to the variable box on your left and click the dependent variable(in this example, it is 'y'.) and move it to 'Dependent:' box by clicking the ▶ button. Similarly, move both independent variables('x1' and 'x2') to 'Independent(s):' box by clicking the ▶ button.

6. Choose Method: Enter . You've got:

7. Click OK in right-upper corner.

8. Now you've got the final results at **Output- SPSS Viewer** window as follows:

Results

Regression

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	X2, X1[a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: Y

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.948[a]	.899	.870	3.446

a. Predictors: (Constant), X2, X1

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	737.276	2	368.638	31.044	.000[a]
	Residual	83.124	7	11.875		
	Total	820.400	9			

a. Predictors: (Constant), X2, X1

b. Dependent Variable: Y

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	2.709	22.358		.121	.907
	X1	.763	.293	.578	2.602	.035
	X2	.463	.251	.409	1.842	.108

a. Dependent Variable: Y

Interpretation

(1) From the last table, estiamted regression coefficients are : β_o = 2.7, β_1 = 0.763, and β_2 = 0.463.

a. Ho: β_1=0

The p-value for testing Ho: β_1=0 is 0.035. We reject the null hypothesis(Ho: β_1=0) and we conclude that sales(Y) is affected by advertising frequency(X1), since p-vlaue = 0.035 < 0.05.

b. Ho: β_2=0

The p-value for testing Ho: β_2=0 is 0.108, which is greater than 0.05. Therefore we do not reject the null hypothesis(Ho: β_2=0) and conclude that sales(Y) is not significantly affected by number of sales persons(X2).

(2) Based on the estimated regression equaition Y= 2.7+0.763*X_1+0.463*X_2, we expect that the sales(Y) will increase by 0.763 units if we increase advertising frequency(X1) by one unit.

(3) R^2 = 0.899 (89.9%) in the second table indicates that the 89.9% of total variance of sales(Y) is explained by the estimated regression equation or, generally, by the advertising frequency(X1) and number of sales persons(X2). Look at the third outputs, ANOVA table.

Source	SS	DF	MS	F	Sig
Regression	SS_R	df_R	MS_R	F	*p-value*
Residual	SS_E	df_E	MS_E		
Total	SS_T	df_T			

$SS_T = SS_R + SS_E$ ➔ 820.4 = 737.28 + 83.12

df_T = number of respondents – 1 = 10 – 1 = 9

df_R = number of independent variables = 2

$df_E = df_T - df_R$ = 9 – 1 = 7

$R^2 = SS_R/SS_T$ = 820.00 / 737.28 = 0.899.

$MS_R = SS_R/df_R$ = 737.28 / 2 = 368.64

$MS_E = SS_E/df_E$ = 83.12 / 7 = 11.87

$F = MS_R/MS_E$ (*Test statistic*) 31.04 = 368.64 / 11.87

→ As F increases, we are more likely to reject the null hypothesis(H_o: all b_j = 0)

P-value associated for this F-statistic is 0.000. Therefore, we conclude that the current regression equation meaningfully describes the relationship between the sales(Y) and advertising frequency(X1) and number of sales persons(X2).