

# Regression Basics for Customer Research<sup>1</sup>

---

This note provides an understanding of the concepts behind regression analysis for customer research. Students might confuse “running a regression using software” with an “understanding of regression.” This note is designed to accomplish the latter.

## Sales and Advertising: A Basic Example

A manager believes increasing spending on marketing-mix factors like advertising leads to higher sales. Statistically, we are interested in knowing if sales and advertising are related. The nature and magnitude of this relationship can be expressed in three different ways, each building upon the other:

- 1) Covariance
- 2) Correlation
- 3) Regression

By now you have probably been exposed to the last two in one form or another. Though it may seem odd, we will also start with the first in our list. The goal of this case exercise will be to:

- 1) Understand the “business context” for the data
- 2) Calculate the covariance between sales and advertising
- 3) Calculate the correlation between sales and advertising
- 4) Calculate the regression line for advertising

Let’s start with the first. Suppose the new advertising director at The Omega Company, Inc. is asked to forecast advertising to be associated with sales. To do this, he is interested in the relationship between advertising and sales. Table 1 shows the data.

Table 1. Sales and Advertising

Year	Sales (\$100,000)	Advertising (\$100,000)
1	10	2
2	15	3
3	20	4
4	25	5
5	30	6
6	35	7
7	40	8
8	45	9
9	50	10
10	55	11
11	60	12
12	65	13
13	70	14
14	75	15
15	80	16
16	85	17
17	90	18
18	95	19
19	100	20
20	105	21
21	110	22
22	115	23
23	120	24
24	125	25
25	130	26
26	135	27
27	140	28
28	145	29
29	150	30
30	155	31
31	160	32
32	165	33
33	170	34
34	175	35
35	180	36
36	185	37
37	190	38
38	195	39
39	200	40
40	205	41
41	210	42
42	215	43
43	220	44
44	225	45
45	230	46
46	235	47
47	240	48
48	245	49
49	250	50
50	255	51
51	260	52
52	265	53
53	270	54
54	275	55
55	280	56
56	285	57
57	290	58
58	295	59
59	300	60
60	305	61
61	310	62
62	315	63
63	320	64
64	325	65
65	330	66
66	335	67
67	340	68
68	345	69
69	350	70
70	355	71
71	360	72
72	365	73
73	370	74
74	375	75
75	380	76
76	385	77
77	390	78
78	395	79
79	400	80
80	405	81
81	410	82
82	415	83
83	420	84
84	425	85
85	430	86
86	435	87
87	440	88
88	445	89
89	450	90
90	455	91
91	460	92
92	465	93
93	470	94
94	475	95
95	480	96
96	485	97
97	490	98
98	495	99
99	500	100

For this case you will:

- 1) Calculate the covariance between sales and advertising

By the end of the case, you should understand the relationship between sales and advertising. You should also understand the relationship between sales and advertising.

Throughout the exercise you will calculate the average and standard deviation of sales. “Transformations” of the sales numbers are used. This will be a good exercise and practice in basic statistics. Each exercise has instructions so you can check your calculations. (Instructions of instructions)

---

<sup>1</sup> © 2018 by Collaborative for Customer-Based Execution and Strategy™. Written by Vikas Mittal, Kyuhong Han, and Jihye Jung. This document is only licensed to be used by permission from The Collaborative for CUBES™. No parts of this case may be copied, reproduced, electronically transmitted, or stored in a retrieval system without permission. For rights and permissions contact: [info@ccubes.net](mailto:info@ccubes.net)

Table 1: Sales vs. Sales and Advertising

Column	Column 1 Sales (100,000)	Column 2 Sales (100,000)	Column 3 Sales (100,000)	Column 4 Sales (100,000)	Column 5 Sales (100,000)
1	100,000	100,000	100,000	100,000	100,000
2	100,000	100,000	100,000	100,000	100,000
3	100,000	100,000	100,000	100,000	100,000
4	100,000	100,000	100,000	100,000	100,000
5	100,000	100,000	100,000	100,000	100,000
6	100,000	100,000	100,000	100,000	100,000
7	100,000	100,000	100,000	100,000	100,000
8	100,000	100,000	100,000	100,000	100,000
9	100,000	100,000	100,000	100,000	100,000
10	100,000	100,000	100,000	100,000	100,000
11	100,000	100,000	100,000	100,000	100,000
12	100,000	100,000	100,000	100,000	100,000
13	100,000	100,000	100,000	100,000	100,000
14	100,000	100,000	100,000	100,000	100,000
15	100,000	100,000	100,000	100,000	100,000
16	100,000	100,000	100,000	100,000	100,000
17	100,000	100,000	100,000	100,000	100,000
18	100,000	100,000	100,000	100,000	100,000
19	100,000	100,000	100,000	100,000	100,000
20	100,000	100,000	100,000	100,000	100,000
21	100,000	100,000	100,000	100,000	100,000
22	100,000	100,000	100,000	100,000	100,000
23	100,000	100,000	100,000	100,000	100,000
24	100,000	100,000	100,000	100,000	100,000
25	100,000	100,000	100,000	100,000	100,000
26	100,000	100,000	100,000	100,000	100,000
27	100,000	100,000	100,000	100,000	100,000
28	100,000	100,000	100,000	100,000	100,000
29	100,000	100,000	100,000	100,000	100,000
30	100,000	100,000	100,000	100,000	100,000
31	100,000	100,000	100,000	100,000	100,000
32	100,000	100,000	100,000	100,000	100,000
33	100,000	100,000	100,000	100,000	100,000
34	100,000	100,000	100,000	100,000	100,000
35	100,000	100,000	100,000	100,000	100,000
36	100,000	100,000	100,000	100,000	100,000
37	100,000	100,000	100,000	100,000	100,000
38	100,000	100,000	100,000	100,000	100,000
39	100,000	100,000	100,000	100,000	100,000
40	100,000	100,000	100,000	100,000	100,000
41	100,000	100,000	100,000	100,000	100,000
42	100,000	100,000	100,000	100,000	100,000
43	100,000	100,000	100,000	100,000	100,000
44	100,000	100,000	100,000	100,000	100,000
45	100,000	100,000	100,000	100,000	100,000
46	100,000	100,000	100,000	100,000	100,000
47	100,000	100,000	100,000	100,000	100,000
48	100,000	100,000	100,000	100,000	100,000
49	100,000	100,000	100,000	100,000	100,000
50	100,000	100,000	100,000	100,000	100,000

Step 1: Sales vs. Sales and Advertising

Based on these data, calculate the correlation and covariance between sales and advertising.

Steps for calculating covariance:  $\sum (X_i - \bar{X})(Y_i - \bar{Y})$  and  $\sum (X_i - \bar{X})^2$  using spreadsheet

- **Step 1** Compute the total and average of Sales and Advertising from the sample data of columns 1 and 2.
- **Step 2** For each row, calculate the "deviation from the mean" for sales as well as advertising.
  - For **column 1**, the deviation is sales from the average (i.e.,  $X_i - \bar{X}$  average sales) = 0
  - For **column 2**, the deviation is advertising from the average (i.e.,  $Y_i - \bar{Y}$ ) = 0
  - **Step 3** The sum for the deviations of 1 should be zero. Same for 2. In other words, columns 1 and 2 should add up to zero.  $\sum (X_i - \bar{X}) = 0$  and  $\sum (Y_i - \bar{Y}) = 0$ .
- **Step 4** Calculate the product of deviations for each row, and write it in Column 3. Multiply the sum of 3 by 2 because of "double-count" or "double-counted".
- **Step 5** Sum up the co-deviations in column 3. Multiply by 40.
- **Step 6** Divide the sum of co-deviations by 40. The answer is covariance that the should be 100,000.

**Directions:** Read the text.

- The number 110,000 has been written in the blank in the number line below.
- The number is 110,000.
- Think: 110,000 is 110 thousand. How will the answer change if all the numbers in column 2 are 110,000 instead?
- How do we address the problem? 110,000 is 110 thousand. How can we use the "units left" to make the number of numbers and place value?

### REVIEW OF BASIC CALCULATIONS

In the space below calculate the standard deviation and variance of sales and advertising.

Please show your calculations by hand in the space below.

Insert the standard deviation and variance of sales and advertising in columns 2 and 3 of Table 1. You will be using 3 data.

**Reminder exercise:** What is the mathematical relationship between variance and standard deviation? In other words, once you know the variance, how can you compute standard deviation and vice versa? Please write your answer in the space below.

## CALCULATING Z-SCORES

Table 2: Calculating Z-scores

Customer Segment	Customer Value	Customer Retention	Customer Lifetime Value	Customer Satisfaction
1	100	80	120	90
2	150	70	110	85
3	200	60	100	80
4	250	50	90	75
5	300	40	80	70
6	350	30	70	65
7	400	20	60	60
8	450	10	50	55
9	500	0	40	50
10	550	0	30	45
Average	300	40	80	70
Standard Dev.	150	20	40	15

- **Step 1:** Calculate the z-score for each entry and fill up columns 4 and 5 of Table 2.  
**Hint:** The z-score is calculated as follows:

$$z = \frac{(X - \mu) / \sigma}{1}$$

**Hint:** These should be inputs of z-scores in column 4 and 5.

- **Step 2:** Calculate, by hand, the average and standard deviation for columns 4 and 5. Show your work.  
**Hint:** The average should be 0, and standard deviation should be 1.

- **Step 3:** Think we have used the same data as variables as "self dependent"? Can you find the advantage of calculating the correlation again, but this time based on z-scores?

**CALCULATING CO-VARIANCE BASED ON Z-SCORES**

Customer	Variable 1 (Z-Score)	Variable 2 (Z-Score)	Variable 1 (Z-Score)	Variable 2 (Z-Score)	Variable 1 (Z-Score)
1					
2					
3					
4					
5					
6					
7					
8					
9					
10					
11					
12					
13					
14					
15					
16					
17					
18					
19					
20					
21					
22					
23					
24					
25					
26					
27					
28					
29					
30					
31					
32					
33					
34					
35					
36					
37					
38					
39					
40					
41					
42					
43					
44					
45					
46					
47					
48					
49					
50					
51					
52					
53					
54					
55					
56					
57					
58					
59					
60					
61					
62					
63					
64					
65					
66					
67					
68					
69					
70					
71					
72					
73					
74					
75					
76					
77					
78					
79					
80					
81					
82					
83					
84					
85					
86					
87					
88					
89					
90					
91					
92					
93					
94					
95					
96					
97					
98					
99					
100					

- **Step 1:** From table 1, record the customer, mean and standard deviations of variables 1 and 2
  - **Step 2:** Calculate the "deviation from the mean" for each variable for customer 1 and 2  
 (Note: these steps are the calculation for variable 1)
  - **Step 3:** In column 3, calculate the co-variance for each variable as the product of columns 4 and 5
  - **Step 4:** Calculate the covariance from these records as the "total" of 3 in column 3
  - **Step 5:** Calculate the covariance based on these records as the "total" of 3 in column 3  
 (Note: use the total and by dividing with 100)
- Step 6:** Consider the results from these records. Are you "lost" the advantage of calculating the co-variance based on z-scores?

**TO BE DONE IN EXCEL or STAT TOOLS for now (But will do it in SPSS also later on)**

Identify significant variables (dependent variable, independent variables, control variables, moderator variables, mediator variables, confounding variables, and interaction variables) and their relationships in a conceptual model. (10%)

Developing a model: Formulate, describe, and represent, and evaluate, an initial model.

1. Identify
2. Identify: nature of dependent and independent variables
3. Identify: relationships of hypothesized relationships of dependent variables

Identify the nature of dependent and independent variables in the model to describe

1. Identifying an independent variable: describe, define, and measure it
2. The dependent variable: describe, define, and measure it. Also, identifying the nature of dependent variable: support the dependent variable to be measured.

**Model**

Model:  $Y = f(X_1, X_2, \dots, X_n) + e$  (where  $Y$  is dependent variable,  $X_1, X_2, \dots, X_n$  are independent variables, and  $e$  is error term)

Model:  $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n + e$

**Model**

Model:  $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n + e$

3. The dependent variable: describe, define, and measure it

Model:  $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n + e$   
 Model:  $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n + e$   
 Model:  $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n + e$   
 Model:  $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n + e$

4. The model:  $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n + e$  and how to interpret the model (10%)

**NOTES (Page left intentionally blank)**



## INTERPRETING REGRESSION OUTPUT FOR CUSTOMER FOCUS: 4-STEP APPROACH

**NOTE:** *Get in the habit of printing out the regression output before you start “interpreting” it*

**Step 1:** Is the overall model good?

**Answer:** Examine the model F-statistic and its statistical significance.

**Step 2:** Which variables / predictors in the model are good?

**Answer:** Use the t-statistic and statistical significance of each variable to answer this question.

**Step 3:** What is the underlying model?

**Answer:** Write the model using the following formulation.

**DV = Intercept +  $\beta$ (Variable 1)\* +  $\beta$ (Variable 2)\* +  $\beta$ (Variable “n”)n.s. ...**

**R<sup>2</sup> = X%**

**MODEL Sig. = p < .05; N=XXXX**

**\* p < .05; n.s. p > .05**

**Step 4:** What is the “interpretation” of the model?

**Answer:** Several points to consider:

- Regression is not reality. It is a representation of reality created via your selection of variables.
- Goal of regression, in customer-focused modeling, is to provide guidance to decision making. Regression is not the decision maker or should not be substitute the decision-making process.
- R<sup>2</sup> provides some understanding of how well your “model” represents reality. However, you can overrepresent reality.

## GUIDELINES IN INTERPRETING REGRESSION FOR CUSTOMER INSIGHTS

1) Building a regression model is a thoughtful process, and ART, not a mechanical science.

**Step 1: Identify the dependent variable and independent variables.**

**Step 2: Check for linearity, normality, and constant variance.**

**Step 3: Interpret the coefficients.**

**Step 4: Evaluate the model fit.**

**Step 5: Use the model for prediction and insight.**

**Step 6: Communicate the findings.**

**Step 7: Validate the model.**

**Step 8: Iterate and refine.**

**Step 9: Document the process.**

**Step 10: Review and update.**

When you have models with many predictors

4) Make it a habit to examine the means and correlations of all variables to get an initial “feel” for how the variables are behaving.

- 1) Given a dataset, a naïve analyst will immediately run 25-30 regressions without knowing what he or she is doing and look to drop non-significant variables
- 2) A smart analyst, will look at the means and correlations, develop an “initial model” and then tweak it. Before, dropping any variable she will ask: what is the meaning of dropping it? Why am I dropping it? Isn’t it important to “control for” its effect? Etc. etc. etc.
- 3) Run at least 4-5 different models to get a “feel” for the underlying story. Which parameters are stable enough that they don’t change no matter if you drop or add variables?
- 4) Do compare the  $R^2$  of different models to see the relative change in model fit. However, sometimes models with lower  $R^2$  are better!

Practice Exercise

Table 14 Customer Satisfaction Survey Data

Customer Number	Overall Satisfaction	Satisfaction with Product

Table 15 Customer Satisfaction Survey Data

Customer Number	Overall Satisfaction	Satisfaction with Customer Service

Compute by hand:

- 1) Covariance of overall satisfaction and product satisfaction (3.9)
- 2) Covariance of overall satisfaction and communication satisfaction (.1)

Compute by hand:

- 1) Covariance of overall satisfaction and product satisfaction (3.9)
- 2) Covariance of overall satisfaction and communication satisfaction (.1)

Overall	Product	Communication	Product	Product	Product	Product	Product	Product
1	1	1	1	1	1	1	1	1
2	1	1	1	1	1	1	1	1
3	1	1	1	1	1	1	1	1
4	1	1	1	1	1	1	1	1
5	1	1	1	1	1	1	1	1
6	1	1	1	1	1	1	1	1
7	1	1	1	1	1	1	1	1
8	1	1	1	1	1	1	1	1
9	1	1	1	1	1	1	1	1
10	1	1	1	1	1	1	1	1
11	1	1	1	1	1	1	1	1
12	1	1	1	1	1	1	1	1
13	1	1	1	1	1	1	1	1
14	1	1	1	1	1	1	1	1
15	1	1	1	1	1	1	1	1
16	1	1	1	1	1	1	1	1
17	1	1	1	1	1	1	1	1
18	1	1	1	1	1	1	1	1
19	1	1	1	1	1	1	1	1
20	1	1	1	1	1	1	1	1
21	1	1	1	1	1	1	1	1
22	1	1	1	1	1	1	1	1
23	1	1	1	1	1	1	1	1
24	1	1	1	1	1	1	1	1
25	1	1	1	1	1	1	1	1
26	1	1	1	1	1	1	1	1
27	1	1	1	1	1	1	1	1
28	1	1	1	1	1	1	1	1
29	1	1	1	1	1	1	1	1
30	1	1	1	1	1	1	1	1
31	1	1	1	1	1	1	1	1
32	1	1	1	1	1	1	1	1
33	1	1	1	1	1	1	1	1
34	1	1	1	1	1	1	1	1
35	1	1	1	1	1	1	1	1
36	1	1	1	1	1	1	1	1
37	1	1	1	1	1	1	1	1
38	1	1	1	1	1	1	1	1
39	1	1	1	1	1	1	1	1
40	1	1	1	1	1	1	1	1
41	1	1	1	1	1	1	1	1
42	1	1	1	1	1	1	1	1
43	1	1	1	1	1	1	1	1
44	1	1	1	1	1	1	1	1
45	1	1	1	1	1	1	1	1
46	1	1	1	1	1	1	1	1
47	1	1	1	1	1	1	1	1
48	1	1	1	1	1	1	1	1
49	1	1	1	1	1	1	1	1
50	1	1	1	1	1	1	1	1

Using a software program that can do the following:

1. Compute using the data matrix
  - a. Total customer, product attributes, and marketing attributes
2. Identify primary needs
  - a. Total customer - all - all products
  - b. Total customer - all - all competitors
  - c. Total customer - all - all products - all competitors
3. Use the results generated to measure the quality of a. The additional insight is you get from using quality 2, which has not yet been