Regression Basics for Customer Research[1]

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

Table 1: Data on Sales and Advertising

Territory	X (SALES)	Y (ADVERTISING)			
1	50	50			
2	100	50			
3	50	50			50
4	50	50			
5	100	50			
Σ					425
Average					
Standard Dev.					
Variance					

Note: X̄ and Ȳ denote average of X and Y, respectively.

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

Steps for calculating covariance (DO BY HAND USING A CALCULATOR, and NOT using spreadsheet):

- Step 1: Compute the total and average of SALES and ADVT (write them in the empty cells of columns 1 and 2).
- Step 2: For each territory, calculate the "deviation from the mean" for sales as well as advertising.
  - For territory 1, the deviation in sales from the average (i.e., X-X̄) is (50 average sales) = 0
  - For territory 2, the deviation in advertising from the average (i.e., Y-Ȳ) = -50
  - Note: The total for the deviations of X should be zero. Same for Y. In other words, columns 4 and 5 should add up to "zero." Make sure you verify this.
- Step 3: Calculate the product of deviations for each territory, and write it in Column 6. Intuitively, this gives you a measure of "co-deviations" or "co-variance"
- Step 4: Sum up the co-deviations in column 6. Verify it is 425.
- Step 5: Divide the total of co-deviations by "n-1". The answer is co-variance (hint this should be 106.25).

[illegible]

- [illegible]
- [illegible]
  - [illegible]
- [illegible]

[illegible]

[illegible]

[illegible]

[illegible]

<u>Reminder exercise</u>: 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

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

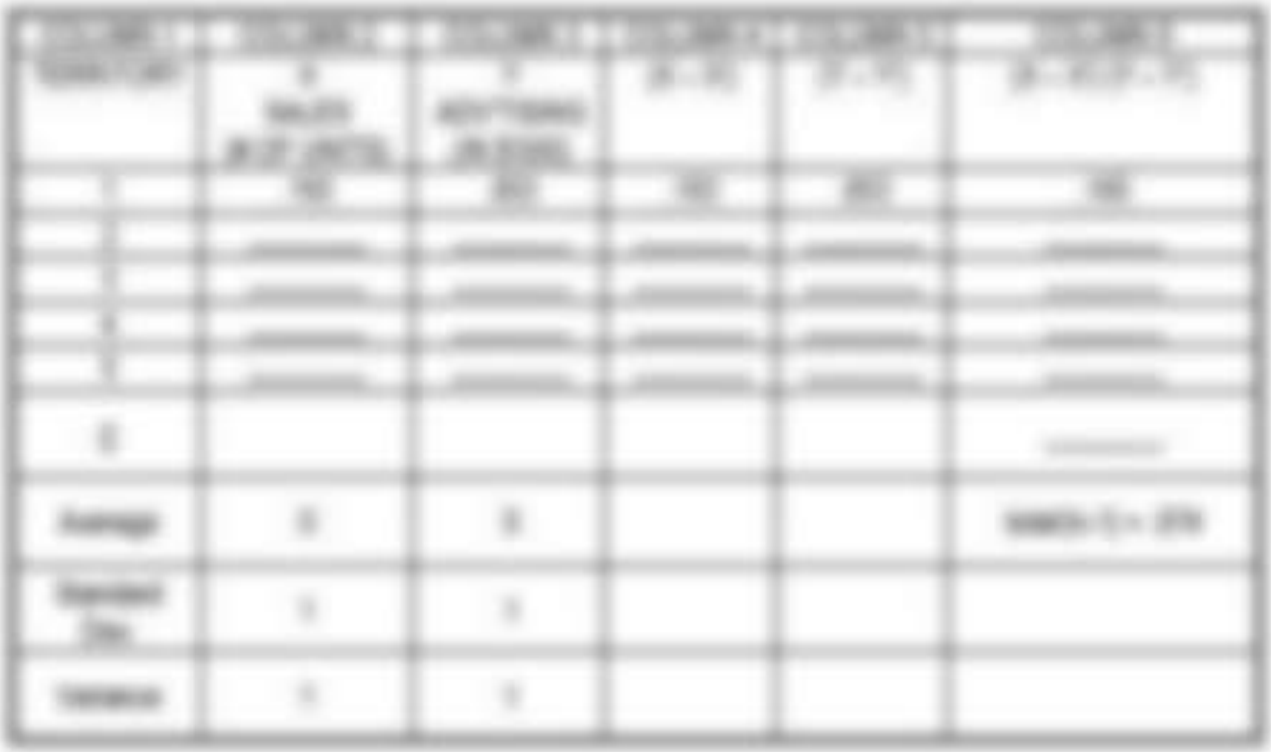
  $z = (X - \bar{X}) \div$ Std. Deviation of X

  Hint: I have shown a couple of z-scores in columns H and I.

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

- Bonus: Earlier we had raised the issue that co-variance is "unit dependent". Can you "test" the advantage of calculating the co-variance again, but this time based on z-scores?

CALCULATING CO-VARIANCE BASED ON Z-SCORES



- Step 1: From table 1, record the z-scores, mean and standard deviations in columns 2 and 3
- Step 2: Calculate the "deviation from the mean" for each territory for columns 4 and 5
  Hint: I have shown the calculation for territory 1

- Step 3: In column 6, calculate the co-variation for each territory as the product of columns 4 and 5
-
- Step 4: Calculate the covariance from these z-scores as the "total/n -1" in column 6
  Hint: I did that and my answer was .874

Bonus: Earlier we had raised the issue that co-variance is "unit dependent". Can you "see" 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)

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 + β(Variable 1)* + β(Variable 2)* + β(Variable "n")$^{n.s.}$ …

R^2 = 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^2 provides some understanding of how well your "model" represents reality. However, you can overrepresent reality.

GUIDELINES IN INTERPRETING REGRESSION FOR CUSTOMER INSIGHTS

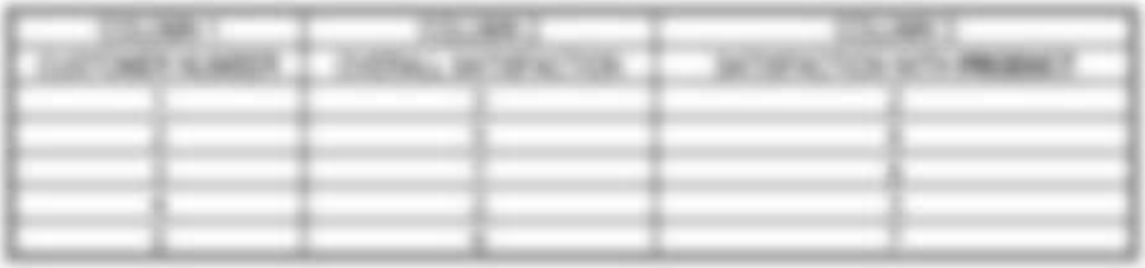
1) Building a regression model is a thoughtful process, and ART, not a mechanical science.

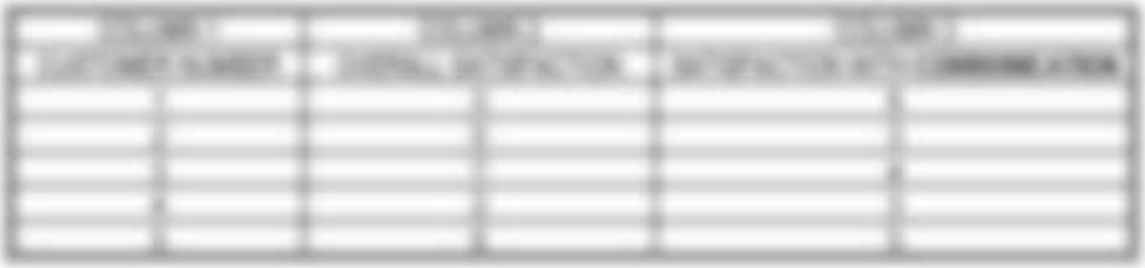
When you have models with many predictors

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





Compute by hand:

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

Using a software program find out the following:

1. Correlation among the three variables
   a. Overall satisfaction, product satisfaction, communication satisfaction

2. Estimate following equations
   a. Overall satisfaction = int. + β(Product)
   b. Overall satisfaction = int. + β(Communication)
   c. Overall satisfaction = int. + β(Product) + β(Communication)

3. Use the 4-step approach to interpret the equation in 2c. What additional insight do you get from using equation 2c, rather than just 2a?