Factor Analysis

Factor analysis is a set of procedures commonly employed for data reduction and summarization, particularly in marketing research where a considerable number of correlated variables need to be streamlined to a manageable level. It involves exploring and representing the relationships among multiple interconnected variables by identifying a few underlying factors. For instance, to assess the image of a fashion brand, participants may be asked to evaluate various competing brands using a semantic differential scale or a Likert scale. The resulting evaluations are then analyzed to reveal the fundamental factors that contribute to the perception of the fashion brand.

Unlike analysis of variance, multiple regression, and discriminant analysis, where one variable is considered dependent or criterion while others are considered independent or predictor variables, factor analysis does not make such a distinction. Instead, it is an interdependence technique that examines the entire set of interrelated relationships. *Factor analysis enables the exploration of possible interconnections between multiple variables and the assessment of the underlying causes behind these relationships.*

Factor analysis finds application in various scenarios:

1. The primary objective of factor analysis is to reveal the underlying dimensions or factors that can explain the correlations among a group of variables. For instance, food habit statements might be employed to gauge consumers' psychographic profiles, which may in turn represent the raw food requirements in particular sectors (and governments/farmers can determine the potential market for it. By subjecting these statements to factor analysis, we can identify the fundamental psychographic factors, as demonstrated in the example given. This is also depicted in Figure 1, which presents the results of empirical analysis indicating that two factors can represent seven psychographic variables. In the figure, factor 1 can be interpreted as the contrast between being a homebody and a socialite, while factor 2 can be seen as the difference between preferences for eating outside and fruits/non-vegan.

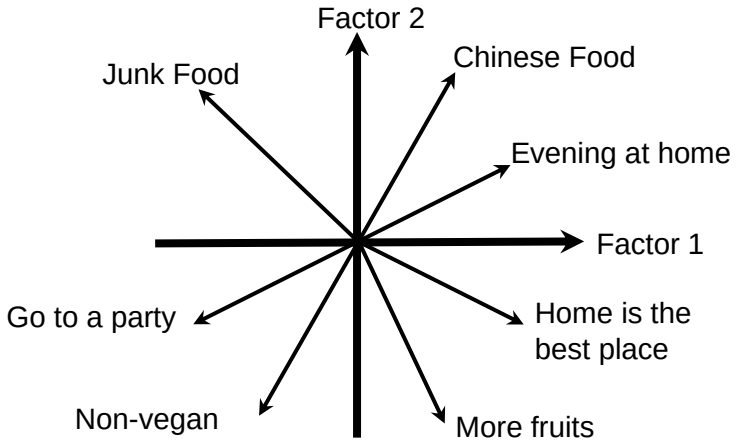

Figure 1

2. Another crucial use of factor analysis is to find a reduced set of uncorrelated variables that can replace the original set of correlated variables in subsequent multivariate analyses, such as regression or discriminant analysis. In the mentioned example, the identified psychographic factors could serve as independent variables to explain the distinctions between loyal and non-loyal consumers. Consequently, instead of

employing the original seven correlated psychographic variables depicted in Figure 1, we can utilize the two uncorrelated factors, namely, homebody versus socialite and vegan versus non-vegan, for further analyses.

3. Factor analysis serves the purpose of identifying a smaller, more meaningful set of variables from a larger set to be utilized in subsequent multivariate analysis. For instance, we can select a few original eating habits statements that strongly correlate with the identified factors and use them as independent variables to explain the distinctions between loyal and non-loyal users.

All these applications of factor analysis are exploratory in nature, earning it the designation of exploratory factor analysis (EFA). Marketing research benefits from numerous uses of factor analysis, such as:

- In market segmentation, it aids in identifying the underlying variables for customer grouping. For example, new tractor buyers may be categorized based on their emphasis on economy, convenience, performance, and comfort, resulting in four segments: economy seekers, convenience seekers, performance seekers, and comfort seekers.

- In product research, factor analysis helps determine the brand attributes influencing consumer choices. For instance, toothpaste brands can be evaluated in terms of attributes like protection against cavities, teeth whiteness, taste, fresh breath, and price.

- In advertising studies, factor analysis can shed light on the media consumption habits of the target market. For instance, users of frozen foods might be found to be heavy viewers of horror films, avid players of electronic games, and listeners of rock music.

- In pricing studies, factor analysis can be employed to identify the characteristics of price-sensitive consumers. For example, such consumers might exhibit traits like being methodical, value-oriented, and home-centric.

Factor Analysis Model

Mathematically, factor analysis bears some resemblance to multiple regression analysis as each variable is represented as a linear combination of underlying factors. The extent to which a variable shares variance with other variables in the analysis is referred to as "communality." The relationships among the variables are characterized by a small number of common factors, along with a unique factor for each variable. These underlying factors are not directly observed. If the variables are standardized, the factor model may be represented as:

$$X_i = A_{i1}F_1 + A_{i2}F_2 + A_{i3}F_3 + \cdots + A_{im}F_m + V_iU_i$$

where X_i $= ith$ standardized variable
A_{ij} = standardized multiple regression coefficient of variable *i* on common factor *j*
F = common factor
V_i = standardized regression coefficient of variable *i* on the unique factor *i*
U_i = the unique factor for variable *i*
m = number of common factors

The unique factors are correlated with each other and with the common factors. The common factors themselves can be expressed as linear combination of the observed variables:

$$F_i = W_{i1}X_1 + W_{i2}X_2 + W_{i3}X_3 + \cdots + W_{ik}X_k$$

where F_i = estimate of i^{th} factor
W_i = weight or factor score coefficient
k = number of variables

To determine these factors, it is possible to select weights or factor score coefficients. The first factor is chosen to account for the largest portion of the total variance. Then, a second set of weights is selected to ensure that the second factor explains most of the remaining variance while being uncorrelated with the first factor. The same principle can be applied to select additional weights for additional factors. This process ensures that the estimated factors' scores, unlike the original variables' values, are not correlated. Additionally, the first factor captures the highest variance in the data, followed by the second factor with the second-highest variance, and so on. Various statistics are associated with factor analysis.

Statistics associated with factor analysis:

The key statistics related to factor analysis include:

- *Bartlett's test of sphericity:* This test statistic examines the hypothesis that variables in population are uncorrelated. In simpler terms, it checks whether population correlation matrix is an identity matrix, meaning each variable has perfect correlation with itself (r = 1) but no correlation with other variables (r = 0).
- *Communality:* Communality refers to the amount of variance a variable shares with all the other variables being considered. It also represents the proportion of variance explained by the common factors.
- *Correlation matrix:* A correlation matrix is a lower triangular matrix displaying the simple correlations (r) between all possible pairs of variables included in the analysis. The diagonal elements, which are all one, are typically omitted.
- *Eigenvalue:* The eigenvalue signifies the total variance explained by each factor.
- *Factor loadings:* Factor loadings are the simple correlations between the variables and the extracted factors.
- *Factor loading plot:* A factor loading plot visually displays the original variables using the factor loadings as coordinates.
- *Factor matrix:* The factor matrix contains the factor loadings of all variables on the extracted factors.
- *Factor scores:* Factor scores are composite scores estimated for each participant based on the derived factors.
- *Factor scores coefficient matrix:* This matrix holds the weights, or factor score coefficients, used to combine the standardized variables to obtain factor scores.
- *Percentage of variance:* This represents the percentage of total variance attributed to each factor.
- *Residuals:* Residuals are the differences between the observed correlations from the input correlation matrix and the estimated correlations from the factor matrix.
- *Scree plot:* A scree plot is a graphical representation of the eigenvalues against the number of factors in the order of extraction.

Conducting factor analysis:

The steps involved in conducting factor analysis are outlined in Figure 2. The process begins with defining the factor analysis problem and identifying the variables to be factor analyzed. A correlation matrix of these variables is then constructed, and a suitable method of factor analysis is chosen. The researcher must decide on the number of factors to be extracted and the

rotation method to be employed. Following this, the rotated factors are interpreted. Depending on the research objectives, factor scores may be calculated, or surrogate variables selected, to represent the factors in subsequent multivariate analysis. Finally, the fit of the factor analysis model is evaluated. Each step is discussed in greater detail in the following subsections.

Formulate the problem:
Formulating the problem involves several tasks. The researcher first identifies the objectives of the factor analysis. The variables to be included in the analysis are specified based on previous research (quantitative or qualitative), theoretical considerations, and the researcher's judgment. It is crucial that the variables are appropriately measured on an interval or ratio scale. The sample size used should be appropriate, with a general guideline of having at least four or five times as many observations (sample size) as there are variables. However, in certain marketing research situations, the sample size might be small, leading to a lower ratio. In such cases, caution should be exercised when interpreting the results.

To illustrate factor analysis, let's consider a scenario where the researcher aims to determine the underlying benefits consumers seek from purchasing toothpaste. The researcher conducted interviews with 30 participants using a seven-point scale to gauge their agreement with statements related to toothpaste benefits (e.g., preventing cavities, freshening breath). The participants were requested to express their level of agreement with the provided statements using a seven-point scale, where 1 stands for "strongly disagree" and 7 represents "strongly agree."

V_1 : It is important to buy a toothpaste that prevents cavities.
V_2 : I Like a toothpaste that gives shiny teeth.
V_3 : A toothpaste should strengthen your gums.
V_4 : I prefer a toothpaste that freshens breath.
V_5 : Prevention of tooth decay should be an important benefit offered by a toothpaste.
V_6 : The most important consideration in buying a toothpaste is attractive teeth.

The obtained data is shown in Table 1. While this illustration involves a small number of observations for clarity, factor analysis is typically conducted on much larger samples. A correlation matrix is then constructed based on these rating data.

Construct the correlation matrix:
The analytical process is based on a correlation matrix of the variables. Insights can be gained from examining this matrix. For factor analysis to yield meaningful results, the variables should be correlated, which is generally the case in practice. If all the correlations between variables are small, factor analysis might not be appropriate. Variables that are highly correlated with each other are expected to have high correlations with the same factor or factors.

Formal statistical tests are available to assess the appropriateness of the factor model. Bartlett's test of sphericity evaluates the null hypothesis that the variables are uncorrelated in the population, meaning the population correlation matrix is an identity matrix. A significant test statistic supports the rejection of the null hypothesis, questioning the suitability of factor analysis.

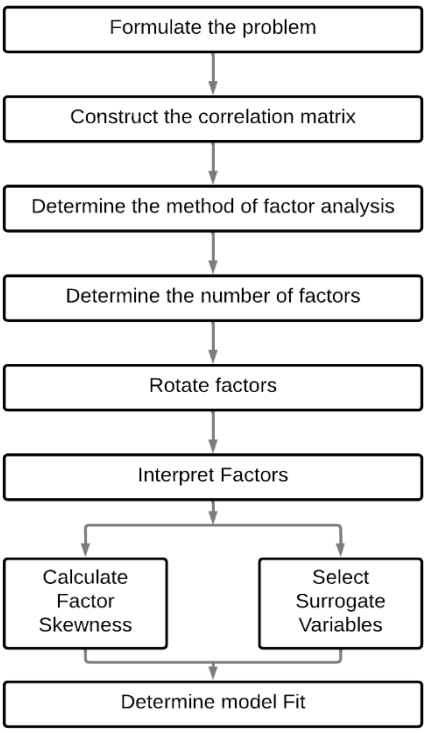

Figure 2

The correlation matrix, constructed from the data obtained to understand toothpaste benefits, are as displayed in Table 2, there are notably strong correlations among V_1 (prevention of cavities), V_3 (strong gums), and V_5 (prevention of tooth decay). It is reasonable to expect these variables to correlate with the same underlying factors. Similarly, V_2 (shiny teeth), V_4 (fresh breath), and V_6 (attractive teeth) exhibit relatively high correlations, suggesting that they might also be associated with the same factors.

The outcomes of the factor analysis can be found in Table 3. Bartlett's test of sphericity rejects the null hypothesis, which assumes that the population correlation matrix is an identity matrix. The calculated chi-square statistic is approximately 111.314 with 15 degrees of freedom, and it is statistically significant at the 0.05 level. Consequently, factor analysis can be considered an appropriate technique for analyzing the correlation matrix shown in Table 2.

Table 1

Participant number	***V_1***	***V_2***	***V_3***	***V_4***	***V_5***	***V_6***
1	7	3	6	4	2	4
2	1	3	2	4	5	4
3	6	2	7	4	1	3
4	4	5	4	6	2	5
5	1	2	2	3	6	2
6	6	3	6	4	2	4
7	5	3	6	3	4	3
8	6	4	7	4	1	4
9	3	4	2	3	6	3
10	2	6	2	6	7	6
11	6	4	7	3	2	3
12	2	3	1	4	5	4
13	7	2	6	4	1	3
14	4	6	4	5	3	6
15	1	3	2	2	6	4
16	6	4	6	3	3	4
17	5	3	6	3	3	4
18	7	3	7	4	1	4
19	2	4	3	3	6	3
20	3	5	3	6	4	6
21	1	3	2	3	5	3
22	5	4	5	4	2	4
23	2	2	1	5	4	4
24	4	6	4	6	4	7
25	6	5	4	2	1	4
26	3	5	4	6	4	7

27	4	4	7	2	2	5
28	3	7	2	6	4	3
29	4	6	3	7	2	7
30	2	3	2	4	7	2

Table 2 a)

Participant	***V_1***	***V_2***	***V_3***	***V_4***	***V_5***	***V_6***
V_1	1.00					
V_2	-0.053	1.00				
V_3	0.873	-0.155	1.00			
V_4	-0.086	0.572	-0.248	1.00		
V_5	-0.858	0.020	-0.778	-0.007	1.00	
V_6	0.004	0.640	-0.018	0.640	-0.136	1.00

Table 2 b)

Factor	***Eigenvalue***	***Percentage of variance***	***Cumulative percentage***
1	2.731	45.520	45.520
2	2.218	36.969	82.488
3	0.442	7.360	89.848
4	0.341	5.688	95.536
5	0.183	3.044	98.580
6	0.085	1.420	100.000

Extraction sums of squared loadings

Table 3 a)

Factor	***Eigen value***	***Percentage of variance***	***Cumulative percentage***
	2.731	*45.520*	*45.520*
2	*2.218*	*36.969*	*82.488*

Factor matrix

Table 3 a)

	Factor 1	***Factor 2***
V_1	*0.928*	*0.253*
V_2	*-0.301*	*0.795*
V_3	*0.936*	*0.131*
V_4	*-0.342*	*0.789*
V_5	*-0.869*	*-0.351*
V_6	*-0.177*	*0.871*

Rotation sums of squared loadings

Table 3 b)

Factor	*Eigenvalue*	*Percentage of variance*	*Cumulative percentage*
	2.688	*44.802*	*44.802*
2	*2.261*	*37.687*	*82.488*

Rotated factor matrix

Table 3 c)

	Factor 1	*Factor 2*
V_1	*0.962*	*-0.027*
V_2	*-0.057*	*0.848*
V_3	*0.934*	*-0.146*
V_4	*-0.098*	*0.854*
V_5	*-0.933*	*-0.084*
V_6	*0.083*	*0.885*

Factor score coefficient matrix

Table 3 d)

	Factor 1	*Factor 2*
V_1	*0.358*	*0.011*
V_2	*-0.001*	*0.375*
V_3	*0.345*	*-0.043*
V_4	*-0.017*	*0.377*
V_5	*-0.350*	*-0.059*
V_6	*0.052*	*0.395*

Reproduced correlation matrix

Table 3 e)

Variables	***V₁***	***V₂***	***V₃***	***V₄***	***V₅***	***V₆***
V_1	*0.926**	*0.024*	*-0.029*	*0.031*	*0.038*	*-0.053*
V_2	*-0,078*	*0.723**	*0.022*	*-0.158*	*0.Q38*	*-0.105*
V_3	*0.902*	*-0.177*	*0.894**	*-0.031*	*0.081*	*0.033*
V_4	*-0.117*	*0.730*	*-0.217*	*0.739**	*-0.027*	*-0.107*
V_5	*-0.895*	*-0.0,8*	*0.859*	*0.020*	*0.878**	*0.016*
V_6	*0.057*	*-0.746*	*-0.051*	*0.748*	*-0.152*	*0.790*

**The lower-left triangle contains the reproduced correlation matrix; the diagonal, the communalities; and the upper-right triangle, the residuals between the observed correlations and the reproduced correlations.*

Selecting the Method of Factor Analysis

Once it has been ascertained that factor analysis is an appropriate methodology for examining the dataset, the next step is to choose the suitable approach. The method employed to establish the weights or factor score coefficients distinguishes the various types of factor analysis. There are two fundamental approaches: principal components analysis and common factor analysis. Principal components analysis takes into account the total variance in the data. In this method, the diagonal of the correlation matrix comprises ones, and the factor matrix encompasses the complete variance. This appro4ach is recommended when the primary objective is to identify the minimum number of factors necessary to explain the maximum variance in the data for subsequent multivariate analysis. These factors are referred to as principal components.

Conversely, common factor analysis estimates factors solely based on the shared variance. Communalities are placed on the diagonal of the correlation matrix. This method is suitable

when the main concern is to recognize the underlying dimensions, and the shared variance is of significance. It is also known as principal axis factoring.

Alternate approaches for estimating common factors also exist, including unweighted least squares, generalized least squares, maximum likelihood, alpha method, and image factoring. These techniques are intricate and are not advisable for inexperienced users. The utilization of principal components analysis on the toothpaste example is illustrated in Table 3.

Determining the Number of Factors

It is possible to calculate as many principal components as there are variables; however, doing so does not lead to simplicity or the revelation of any underlying structure. To summarize the information inherent in the original variables, a smaller set of factors should be extracted. The question then becomes: how many factors should be chosen? Several methods have been proposed for determining the appropriate number of factors. These include a priori determination, approaches based on eigenvalues, the scree plot, percentage of variance explained, split-half reliability, and significance tests:

1. *A priori determination:* Researchers sometimes have prior knowledge that allows them to anticipate the number of factors, enabling them to predefine this number before factor extraction. Most software applications facilitate this approach by permitting users to specify the desired number of factors.
2. *Eigenvalue-based determination:* This approach retains only factors with eigenvalues surpassing 1.0, discarding the remaining factors. An eigenvalue signifies the amount of variance linked to a factor. Consequently, only factors with variances greater than 1.0 are retained, as factors with variances below 1.0 are no more informative than individual variables, each of which has a variance of 1.0. For datasets with fewer than 20 variables, this method generally yields a conservative number of factors.
3. *Scree plot-based determination:* A scree plot displays eigenvalues against the order of factor extraction. The plot's shape guides the decision on the number of factors. Typically, the plot exhibits a distinct break between the steep decline of eigenvalues corresponding to significant factors and a gradual decline for the remaining factors. This gradual descent is referred to as the "scree." Empirical evidence suggests that the point where the scree begins marks the true number of factors. Normally, the number of factors derived from the scree plot will be slightly higher than that obtained from the eigenvalue criterion.
4. *Percentage of variance-based determination:* This approach selects the number of factors in a manner that ensures the cumulative percentage of variance explained by the factors reaches an acceptable level. The definition of an acceptable level depends on the specific problem. It is generally advised that the extracted factors account for at least 60% of the variance.
5. *Split-half reliability-based determination:* The dataset is divided into two halves, and factor analysis is conducted on each subset. Only factors with substantial agreement in factor loadings across the two subsets are retained.
6. *Significance tests-based determination:* This approach involves assessing the statistical significance of individual eigenvalues and retaining only those factors with statistically significant eigenvalues. A limitation of this approach is that in large samples (size exceeding 200), numerous factors may be statistically significant, yet many of these contribute only minimally to the overall variance.

Table 3 presents the application of the eigenvalue criterion, resulting in the extraction of two factors. Prior knowledge suggests that toothpaste purchases are driven by two primary reasons. The corresponding scree plot can be observed in Figure 3, indicating a clear break at three

factors. Furthermore, by considering the cumulative percentage of variance accounted for, it becomes apparent that the first two factors explain 82.49% of the variance, and progressing to three factors offers marginal improvement. Additionally, the split-half reliability analysis supports the appropriateness of two factors. In conclusion, this situation seems to justify the selection of two factors.

The "extraction" column within the "Communalities" section of Table 3 furnishes pertinent details after the desired number of factors has been extracted. Notably, the communalities under "Extraction" differ from those under "Initial" because the variances attributed to the variables remain unexplained unless all factors are retained. The Table 3a labelled "Extraction sums of squared loadings" provides the variances linked to the retained factors.

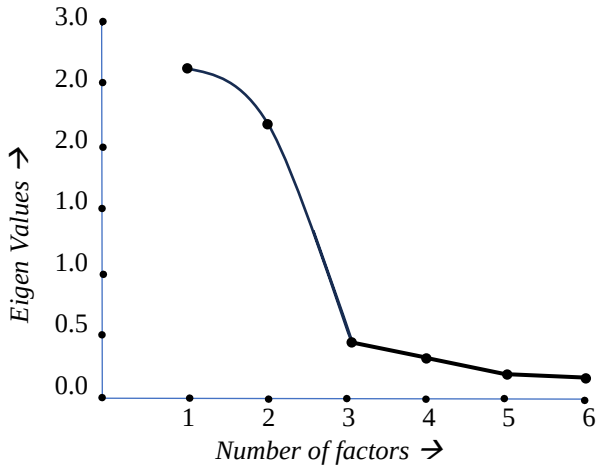

Figure 3

Multivariate Multiple Linear Regression

Multivariate multiple linear regression is a way to model the linear link between more than one independent variable and more than one dependent variable. It has more than one independent variable, so it is multiple, and it has more than one dependent variable, so it is a multivariate model.

Assumptions Underlying Multivariate Multiple Linear Regression

Assumptions are inherent to every statistical technique. Assumptions signify the prerequisite properties that your data must possess to ensure the accuracy of statistical method outcomes.
The assumptions pertinent to Multivariate Multiple Linear Regression encompass the following aspects:

1. *Linearity:* The variables of interest must exhibit a linear relationship. This implies that when plotting these variables, a straight line should effectively capture the data's shape.
2. *Absence of Outliers:* The variables under consideration should not contain outliers. Linear Regression is sensitive to outliers—data points with exceptionally large or small values. Identifying outliers involves plotting the variables and identifying points that substantially deviate from the majority of other points.
3. *Homoscedasticity*: Also known as similar spread across the range, homoscedasticity signifies that variables maintain consistent dispersion across their respective ranges.
4. *Normality of Residuals:* "Residuals" pertain to the discrepancies between expected (or predicted) dependent variable values and the actual values. These discrepancies' distribution should conform to a normal (bell curve) distribution shape. Meeting this assumption ensures that the regression results are equally valid across the data's entire spread, devoid of any systematic bias.
5. *No Multicollinearity:* Multicollinearity occurs when two or more independent variables demonstrate significant correlation among themselves. This situation renders regression coefficients and statistical significance unstable and less reliable. However, it doesn't inherently impact the model's goodness of fit.

Employ Multivariate Multiple Linear Regression in the following scenarios:

1. *Prediction:* When seeking a statistical tool to predict one variable using another, the situation calls for a prediction-oriented analysis. Other types of analyses involve examining the strength of the relationship between two variables (correlation) or comparing differences between groups (difference).
2. *Continuous Dependent Variable:* The variable you intend to predict must be continuous. Continuous variables can assume a broad range of values, such as heart rate, height, weight, etc. Data types like ordered data, categorical data, or binary data are not continuous.
3. *Multiple Independent Variables:* Multivariate Multiple Linear Regression is suitable when one or more predictor variables have multiple values for each unit of observation.
4. *No Repeated Measures:* This method applies when there's only one observation for each unit of observation. Units of observation constitute individual data points, such as a store, customer, city, etc. For cases with repeated measurements from the same group over time, a Mixed Effects Model should be used.
5. *More than One Dependent Variable:* To utilize Multivariate Multiple Linear Regression, you should have multiple dependent variables or variables you are aiming to predict. For a single dependent variable, Multiple Linear Regression suffices.

Example 1:
In a dairy business, a farmer association plans to distribute milk across cities and they have assigned a budget on advertising at different cities and also recorded the city population. The association wishes to determine the revenue generated and customer traffic.

Approach to solution:
Dependent Variable 1: Revenue
Dependent Variable 2: Customer traffic
Independent Variable 1: Advertising expenditure by city
Independent Variable 2: City Population

The null hypothesis posits that there is no relationship between advertising spending and revenue or population by city. Our analysis assesses the validity of this hypothesis.

After ensuring adherence to linear regression assumptions, the analysis is performed. This entails conducting multiple linear regressions for each dependent variable. Consequently, beta coefficients and p-values are derived for both the "revenue" and "customer traffic" models. Each linear regression model includes an intercept beta coefficient (β_0) and potentially additional beta coefficients (β_1, β_2, etc.) representing the relationships between independent and dependent variables.

These additional beta coefficients offer insights into the quantitative relationship between variables. A unit increase in a given independent variable corresponds to a change in the dependent variable by the value of the associated beta coefficient (with other independent variables held constant).

The p-value linked to these beta values represents the probability of observing the results under the assumption of no actual relationship between that variable and revenue. A p-value ≤ 0.05 indicates statistical significance, implying that the difference is not due to chance alone. Overall p-values for the model and individual p-values representing variable effects can be obtained through Multivariate Analysis of Variance (MANOVA).

Moreover, the analysis yields an R-Squared (R2) value, ranging from 0 to 1, denoting the goodness of fit between the linear regression line and data points. A higher R2 signifies better model fit.

Example 2 (for practice):
A researcher wishes to determine what factors affect the health of Sunflower plants. He accumulates information on the dependent variables such as average leaf diameter, the mass of the root ball, and the average bloom diameter, as well as the length of time the plant has been in its current container. He measures several soil elements, as well as the quantity of light and water each plant receives, as independent variables.

References:

1. Malhotra, N., Nunan, D. and Birks, D., 2017. Marketing Research: An Applied Approach. Pearson.

2. Yeater, K.M. and Villamil, M.B., 2018. Multivariate methods for agricultural research. Applied statistics in agricultural, biological, and environmental sciences, pp.371-399.

3. Goddard, W. and Melville, S., 2004. Research Methodology: An Introduction. Juta and Company Ltd.

4. Nesselroade, J.R. and Cattell, R.B. eds., 2013. Handbook of Multivariate Experimental Psychology. Springer Science & Business Media.

5. https://www.statstest.com/multivariate-multiple-linear-regression/

6. https://www.voxco.com/blog/multivariate-regression-definition-example-and-steps/

7. https://stats.oarc.ucla.edu/stata/dae/multivariate-regression-analysis/