Regression Model Building with MS Excel: Using Excel's Multiple Regression Tool to Explore the Correlation Between Advertising Dollars and Sales Volume

Andy Bertsch
Minot State University

Alisa Dahl
Minot State University

Using a common set of indicators, this paper explores useful concepts such as correlation, error, confidence intervals, and quantitative equation modeling using MS Excel's Multiple Regression tool. These concepts, provided and explained in a straightforward manner, coupled with the common tools within MS Excel, may allow managers greater understanding of the risk and uncertainty involved in many daily decisions.

INTRODUCTION

Advertising in America traces its roots back to the 1700's when Benjamin Franklin's Pennsylvania Gazette included pages of advertisements. Since that time, advertising has evolved to the point where it is difficult for the industry to agree on a definition.

No matter how one defines "advertising", most would agree advertising is designed to produce a response – an increased desire for the product on the part of the consumer, resulting in purchases which ultimately bring about an increase in sales. Because of this, it is often assumed that an increase in the amount a company spends on advertising will automatically bring about an increase in sales. Is this a correct assumption? Is there a relationship between advertising expense and sales? This seeks to answer: What affect does advertising expense have on sales when other variables such as changes in population, unemployment and number of competitors are present?

LITERATURE REVIEW

Advertising has a rich history in America beginning in the 18th century. As new forms of media were introduced, advertising expenditures continued to grow and reach larger target populations. By the beginning of the 20th century, marketers were spending double the amount

on newspaper and magazine advertising compared to what they had 20 years earlier in an attempt to create demand for their products, and it worked (Rothenberg, 2005).

One of the biggest boosts to advertising came with the invention of television, which made an immediate impact. Rothenberg (2005) writes in his essay for The Advertising Century, "Hazel Bishop lipstick sales skyrocketed from $50,000 a year in 1950 to $4.5 million two years later thanks to TV advertising." In this case expenditures for advertising brought about the desired result of increased sales.

Other evidence has been presented to support the idea that successful advertising campaigns are linked to increases in sales. Research was conducted in 1965 to determine whether advertising had a measurable effect on purchases of Oscar Mayer wieners (Twedt, 1965). Results indicated advertising had a measurable effect on brand purchase – each advertising dollar spent resulted in about $2.50 in additional sales. At that time in history it was believed each dollar of advertising spent resulted in an average of $3.00 in additional purchases (Twedt, 1965).

In an attempt to determine how to effectively allocate marketing resources over the life span of a new product, a model of consumer response to new products was developed in 1973 to determine a method of forecasting new product sales by measuring changes in marketing activities (Nakanishi, 1973). Results indicated advertising and promotion effects on consumer response to new products are different depending upon the medium used. In addition, advertising expenditures in the form of network TV commercials, local spot TV commercials, and direct mail coupons for a new product resulted in an increase in purchases of the product, with direct mail coupons having the largest results.

There is evidence that advertising does not always lead to an increase in consumer desire for the product. In the case of New Coke, an increase in advertising to create a campaign designed to promote the new product did not result in an increase in sales. In this campaign, advertising attempted unsuccessfully to change the image of the product. There was no correlation between the amount spent on advertising and sales – the new formula taste was not what hardcore Coke consumers wanted resulting in a decrease in sales. Another example of an unsuccessful marketing campaign is Snapple. Once marketed by Quaker Oats Company, the soft drink could not penetrate the market successfully and the company was finally sold in 1997 for the low price of $300 million. The new owner recognized a problem not with the product but with the advertising campaign and was successful in stabilizing the brand (Lewis, Nelson, 1999).

Although there are always exceptions, it is generally believed in most cases advertising does have an impact on increasing desire. Because of this, advertising expenditures are essential to promoting sales. At one time, Proctor & Gamble believed that everyone knew the advantages of Ivory soap so they slashed their advertising budget to zero. This resulted in their sales plummeting to an all-time low (Lewis et al, 1999). This would seem to indicate a minimum advertising budget is required to maintain sales.

Today, most American advertisers still believe the more you spend the better the results. In fact, major advertisers spend as much on their advertising campaigns as they earn from their brands (Jones, 2000). However, many companies today are disappointed in the results of their advertising investment because only about a third of all ad campaigns have a significant immediate impact on sales - less than one fourth have any long-term effect. In his article "The Mismanagement of Advertising", Jones (2000) speculates the problem is a result of advertising being under-emphasized by management resulting in some great ads going by the wayside and bad ideas being implemented.

In his book Why We Buy, Paco Underhill (2000) explains that branding and traditional advertising help to build brand awareness and a predisposition towards purchasing, but unfortunately this do not always mean an increase in sales. He believes the standard tools of marketing and advertising still work, they just do not work as well as they used to. Developing an advertising campaign is not enough to increase sales; signage, shelf positions display space, store fixtures, and the like become just as important. This phenomenon has helped to create the science of shopping. Advertising may get the consumer in the door but the science of shopping is what determines if and what they will buy once they get there.

Advertising dollars spent to promote products via the Internet have proven to bring about desired results. For Ford Motor Co., a promotion launched simultaneously on leading portals exposed some 50 million Web surfers to the new Ford banner. Millions of surfers visited Ford's site at a rate of 3,000 per second resulting in a 6% jump in sales over the first three months of the campaign (Baker, 2004). Ford has dedicated 10% of its advertising budget to online advertising along with most other auto makers. Internet ads are now considered mainstream and part of the advertising expenditures because they bring about the desired results in the mass market arena.

Historically, the determination of how much to spend on advertising has predominately been the percentage of sales method (based on sales of the product). Chakrabarti's research published in 2007 indicates the gross profit margin before marketing is the single most important element in determining advertising to sales ratios.

Other factors may influence the impact that advertising dollars have on sales. Such factors may include population, competition, and unemployment rates. Intuitively, areas with greater population ought to result in higher sales for each incremental increase in advertising dollars. In addition, other research has linked the number of competitors to sales income; sales response to advertising could be higher if there is less competition (Danaher, et al., 2008). Finally, slow sales growth is usually expected when unemployment rises. Recent articles seem to support this expectation (For Most Retailers, Road Looks Rough, 2008; Economists Say Recession Here, 2008). It is anticipated these variables will also influence what affect advertising has on sales income.

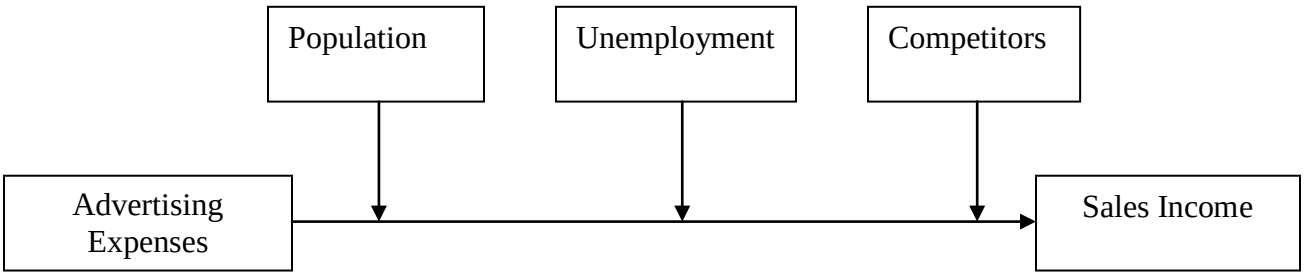
HYPOTHESIS

The literature discussed the evolution of advertising throughout the years. In the early age of advertising, spending money on advertising was all that was required to bring about desired sales. Today, many other factors come into play as outlined in the Paco Underhill's book "Why We Buy: The Science of Shopping". No matter what other influences cause consumers to buy a product, literature reveals that advertising expenditures are required to promote sales, albeit exceptions exist when advertising campaigns are faulty or based on false premises. In addition, this paper will examine whether other factors such as unemployment, population, and number of competitors will influence the affect advertising has on sales income. For this reason, the hypothesis for this study is as follows:

> **Hypothesis**: Sales income is dependent upon advertising expenses, regardless of other variables such as unemployment, population, and number of competitors.

The null hypothesis implies that the 95% confidence interval for the 'other' regression coefficients will contain zero (0) and, consequently, will be of no use in estimating the dependent variable (sales income).

FIGURE 1
HYPOTHESIS



METHODOLOGY

In order to build and test an appropriate model for this study, it was necessary to locate a relevant dataset. The dataset used to build and test the model for this study was borrowed from Harnett & Horrell (1998). The information consisted of sales information for 'Ohio Valley Detergent' over the course of 30 quarters. Additional information in the dataset covering the same quarters included the local unemployment rate, the population in the metropolitan area, the amount spent on advertising, and the number of competitors in the local market.

TABLE 1
DATASET VARIABLES

Sales Income	Dependent Variable
Unemployment Rate	Independent Variable 1
Population	Independent Variable 2
Advertising Expenses	Independent Variable 3

Descriptive analysis was completed on the independent variables listed in Table 1 above. This analysis provided information such as the mean, median, mode, standard deviation, and range of the data being analyzed. In order to determine whether a linear relationship is present between sales income and the variables listed above, scatter diagrams were plotted showing the dependent variable (sales income) and each independent variable separately. These diagrams help to visually determine whether a linear relationship exists between the dependent and independent variables indicating the direction, linearity, and strength of each relationship (independently).

Because simple linear regression is limited to one variable at a time and does not demonstrate possible interaction among independent variables, multiple regression analysis was also completed in order to conduct a test of the hypothesis on each of the regression coefficients. Multiple regression analysis included eliminating the variables which have confidence intervals that straddle zero (Harnett & Horrell, 1998). If a slope coefficient is zero (0), then regardless of the value of that independent variable, the dependent variable is constant. Therefore, a slope coefficient of zero (0) indicates that that particular variable will have no effect on the dependent variable in the resulting model. If zero (0) is contained in the confidence interval, the variable is not useful in the resulting model. These are considered 'bad' variables (Harnett & Horrell, 1998; Lind, et al., 2008) and need to be removed in an iterative process (one at a time). Once the "bad" variables were eliminated, the best model was selected by examining the standard error of the

resulting model while retaining as many variables as possible. Finally, the regression equation was determined from the data (in order to predict sales income based on the independent variables remaining in the final regression model), as well as upper and lower boundaries for confidence intervals, the standard error of the estimate, and the coefficient of determination in the final multiple determination (R^2). These figures will help to determine the relationship between the dependent and independent variables.

RESULTS

Descriptive Statistics

Data developed in the Descriptive Analysis table was reviewed for the sample of 30 quarters of the Detergent Sales dataset. This information provides a better understanding of the data being analyzed. For instance, it is known that during the sample period of 30 quarters, Sales Income ranged from a high of $12,800,000 to a low of $5,100,000 with an average of $7,380,330. During that same period, Advertising Expenses ranged from a high of $78,900 to a low of $25,000 with an average spent per quarter of $55,896.67. Summaries of the data are shown in Table 2 below:

TABLE 2
SUMMARY OF DESCRIPTIVE ANALYSIS

	SALES INCOME (in 000s)	ADVERTISING EXPENSE	POPULATION (in 000s)	UNEMPLOYMENT RATE	COMPETITORS
MEAN	7380.33	55896.67	1169.8	5.81	2.53
STANDARD ERROR	363.41	2877.23	23.09	.09	.27
MEDIAN	6750	53100	1140	5.7	2.5
MODE	6200	45000	1100	5.4	1
STANDARD DEVIATION	1990.45	15759.21	126.47	.49	1.46
RANGE	7700	53900	393	1.8	5
MINIMUM	5100	25000	987	5	0
MAXIMUM	12800	78900	1380	6.8	5

Linear Relationships

Scatter diagrams were created which plotted the dependent variable against each independent variable in order to visualize the relationships between the variables and indicate such attributes as strength, direction, and the like. Results indicated all variables demonstrated a linear relationship. The strongest positive correlation was found in the variable Population, which indicated 59.4% of the variation in Sales Income can be explained by the variation in population (r=0.771). The Unemployment Rate demonstrated the weakest correlation (negative). The

negative correlation is intuitively logical as one would assume a rise in unemployment rate would bring about a decrease in sales. Correlation Coefficients (r) and Determination Coefficients (R^2) are listed in the following table.

TABLE 3
LINEAR RELATIONSHIPS

VARIABLE	DIRECTION	CORRELATION COEFFICIENT	DETERMINATION COEFFICIENT
Advertising Expense	Positive	0.440112	0.193698
Competition	Positive	0.662979	0.439541
Population	Positive	0.771	0.59441
Unemployment Rate	Negative	-0.320528	0.102738

Multiple Regression Analysis

In the event there are "bad" variables, iterative runs of multiple regression are necessary to ferret out those variables that are, indeed (and independently) "bad". Multicollinearity can give the impression that a variable is disruptive simply because it is collinear with a "bad" variable. At times and depending on the number of independent variables, the number of iterative runs of multiple regression (removing one variable at a time and adding back in previously suspicious variables) can be time consuming and frustrating. But it is a necessary process in order to determine the best possible multiple regression model (Harnett & Horrell, 1998; Lind, et al., 2008).

The first run of the multiple regression analysis revealed two variables which had confidence intervals crossing zero: Unemployment Rate and Competition. The second run eliminated Unemployment Rate and the results indicated Competition still straddled zero for confidence intervals. The third run eliminated Competition and added back in Unemployment Rate. Results indicated Unemployment Rate still had zero as a possibility in its confidence intervals. Because of this, both Unemployment Rate and Competition were eliminated from the final regression analysis table. Final results are outlined below in Table 4.

TABLE 4
MULTIPLE REGRESSION – DETERGENT SALES

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.809494343
R Square	0.655281091
Adjusted R Square	0.629746357
Standard Error	1211.157611
Observations	30

	Coefficients	Standard Error	T Stat	P-value	Lower 95%	Upper 95%
Intercept	-7385.913459	2108.36258	-3.503151464	0.0016203	-11711.91324	-3059.91
Population (in 000s)	11.08033192	1.842801835	6.012763665	2.0457E-06	7.299217409	14.86145
Advertising Expense	0.032282328	0.0147884	2.18294927	0.03790933	0.001939057	0.062626

The results above were used to develop the following linear equation:
SALES INCOME = -7385.9135 + 11.0803 (population) + .0323 (advertising expense)

Using this equation, an estimate of sales based on changes in the independent variables is possible. For example, increasing Advertising Expense from the $53,900 spent in the final quarter in the dataset to $80,000, a reasonable prediction can be made for Sales Income assuming the population remains the same.
SALES INCOME = -7385.9135 + 11.0303(1340) + .0323 ($80,000)
SALES INCOME = $10,045.69 thousand or $10,045,690

The regression equation for this dataset indicates for every $1000 increase in advertising expenses, there will be a $32,000 increase in sales income. Every 1000 increase in population brings about an $11,000 increase in sales income.

Confidence Intervals

Our regression analysis in Table 4 provides information to calculate the upper and lower boundaries at a confidence level of 95% for the sample equation listed above.
Upper boundary = -3059.9137 + 14.8614(1340) + .0626($80,000) = $21,862,361.30
Lower boundary = 0 + 7.2992(1340) + .0626($80,000) = $9,932,928

In order to find a more accurate prediction of confidence intervals, a more accurate method is available using the degrees of freedom and the model's standard error. The resulting formula and resulting confidence intervals are:

C.I. = Sampled sales income +/- t(standard error)
C.I. = 10045.6885 +/- 2.05182914 * 1211.157611

Upper boundary = 12530.77698 = $12,530,776.98
Lower boundary = 7560.600021 = $7,560,600.02

As can be seen, this is a much tighter interval. This is a better method to use because it takes the total error of the model into consideration. This demonstrates that sales could range between $12,530,776 and $7,560,600 and still be within the 95% confidence interval.

Multiple Standard Error of Estimate

The standard error of estimate is comparable to the standard deviation and utilizes squared deviations from the regression line. In this multiple regression equation, the standard error of estimate is 1211.16, as listed in Table 4. This would mean one could expect 68% of the residuals to be within ± $1,211,000 and 95% to be within ± $2,422,000. This may appear to be a large variation, but sales income actually ranges from $5,100,000 to $12,800,000 in the given dataset. The reader is cautioned that there may also be other variables that would make the model more robust. For the purposes of this particular study, the variables used in this model were limited to those readily available in the selected dataset.

Coefficient of Multiple Determination

The Coefficient of Multiple Determination (R^2) is the percent of variation in the dependent variable Y explained by the resulting collective set of independent variables Advertising Expense and Population (Excel labels this "R squared"). The result of 0.65528 shown as "R Squared" in Table 4 tells us that 65.6% of the variability in sales income is explained by the variables of Advertising Expense and Population. This is a fairly good model as only 35% of the total variance in Y is unexplained by these two variables. On the other hand, one may wonder what the other variables are that will explain part of the remaining 35% of the variance in Y. When one considers the "adjusted" coefficient of determination (listed as "Adjusted R Squares" in Table 4), a more accurate picture of the model's goodness of fit is approximated. This is true because the adjusted R-squared take into account the total number of independent variables in the model (Harnett & Horrell, 1998; Lind, et al., 2008). With only two variables, the "Adjusted R Squared" is a bit lower at 63%. Such a value of R-squared, although significant, illustrates that other variables may be at play as was previously suggested.

CONCLUSIONS

As theorized at the beginning of this report, the data supports the hypothesis that advertising expenses are related to sales, regardless of other variables present. Although the literature seemed to indicate a strong relationship existed, the results from the dataset studied support the fact that other variables are also significant. In addition, the regression analysis for the dataset did not support the literature which showed that the number of Competitors and Unemployment Rates influence Sales Income. The analysis was conducted using MS Excel's Multiple Regression tool.

The linear relationship between Advertising Expenses and Sales Income as demonstrated on a scatter diagram is weak with a 19.4% Determination Coefficient. For this dataset, only 19.4% of the variation in Sales Income can be explained by the variation in Advertising Expenses alone. The linear relationship between number of competitors and Sales Income as demonstrated on a scatter diagram is moderate with a 43.9% Determination Coefficient. This would indicate that

44% of the variation in Sales Income can be attributed to the variation in the number of competitors. However, after running the regression analysis, variables were eliminated because coefficient confidence intervals straddled zero. This was very perplexing and begs further research using additional datasets since research suggests number of competitors do have influence on sales outcome.

The linear relationship between the Unemployment Rate and Sales Income supported the elimination of this variable in the regression analysis since the relationship was very weak. Finally, the linear relationship between Population and Sales Income was stronger than all other variables and the regression analysis supported the use of this variable for inclusion in the regression equation. This is logical since a larger population should bring about higher sales.

The analysis indicated that for this company, every $1,000 increase in advertising expenditures should result in a $32,000 increase in Sales Income. It also demonstrated that one can expect an $11,000 increase in Sales Income for every 1,000 increase in Population.

Based on the information presented in the scatter diagrams and regression analysis, the hypothesis should be accepted. However, the fact that the relationship between Advertising Expense and Sales Income was not as strong as expected seems to support recent literature which indicates consumers are being bombarded with advertising to the point that the "clutter" impacts the positive effects on Sales Income (Danaher, et al., 2008).

REFERENCES

Baker, S. (2004, November 22). The Online AD SURGE. Business Week.

Chakrabarti, S. & Arora, A. (2007). Drivers of ad spend trends among global marketers. International Journal of Advertising, 26(3), 387-397.

Danaher, P., Bonfrer, A. and Dhar, S. (2008, May). The Effect of Competitive Advertising Interference on Sales for Packaged Goods. Journal of Marketing Research (JMR), 45(2), 211-225.

Economists Say Recession Here (2008, March 17). Money Management Executive.

For Most Retailers, Road Looks Rough (2008, February 25). Dow Theory Forecasts.

Jones, J. (2000, January). The Mismanagement of Advertising. Harvard Business Review, 78(1), 22-23.

Harnett, D. L. & Horrell, J. F. (1998) Data, Statistics, and Decision Models. New York: Wiley.

Lind, D. A., Marchal, W. G., Wathen, S. A. (2008) Statistical Techniques in Business & Economics, New York: McGraw Hill Irwin

Lewis, H. & Nelson, C. (1999). Advertising Age: Handbook of Advertising. Lincolnwood (Chicago), IL: NTC Business Books.

Nakanishi, M. (1973, August). Advertising and Promotion Effects on Consumer Response to New Products. Journal of Marketing Research (JMR), 10(3), 242-249.

Rothenberg, R. (2005). The Advertising Century, AdAge.Com, retrieved February 20, 2008 from http://adage.com/century/rothenberg.html.

Twedt, D. (1965, April). How Can the Advertising Dollar Work Harder? Journal of Marketing, 29(2), 60-62.

Underhill, P. (2000). Why We Buy: The Science of Shopping. New York, NY: Simon & Schuster Paperbacks.