

# Geographic Determination of a Successful ACO Hardware Store

Michael Sabor  
Ana Rita Silva  
Matthew St. Peter

December 13, 2006

**Abstract** ACO Hardware has gathered a large amount of diverse and conflicting financial data, combined with site, demographic, traffic, consumer behavior and competition data in an effort to build a "successful store profile". This profile would then be used as a tool for evaluation when developing a new location and expanding or moving an underperforming location. The focus of this project is to evaluate ACO Hardware's current operations, extract the relevant data, and develop a ten-factor model to evaluate success.

## Table of Contents

Introduction .....	1
Multiple Linear Regression .....	2
Data Processing .....	2
Data and Residual Analysis .....	4
Discussion of the Factors .....	7
Conclusion .....	9
Future Work .....	9
Acknowledgements .....	9
References .....	10
Appendix .....	11

## **Introduction**

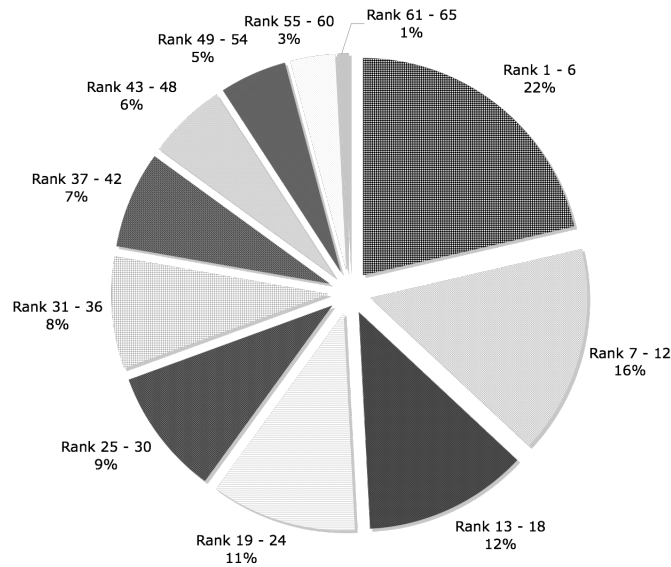
ACO currently operates 69 retail locations throughout Michigan, primarily situated in the metro Detroit area, though stores span as far east as Battle Creek, MI and as far north as Bay City, MI. Like many retailers, the company is in a transitional period. The expansion of “big box” hardware stores such as Home Depot and Lowe’s has created a new burden to remain competitive. In the face of this challenge, ACO has chosen to occupy a smaller niche, preferring to cater to the small home improvement market rather than compete directly with the larger retailers.

As part of their plan to remain a strong retail operation, ACO is in the process of a multi-faceted business review process, attempting to identify the factors that will contribute to continued success and expansion. ACO has gathered a large amount of financial data, combined with site, demographic, traffic, consumer behavior and competition data, in an effort to build a “successful store profile” that may be used as a basis for expansion.

However, ACO has had difficulties interpreting the gathered data. The company has primarily used Microsoft Excel to compare data, which is insufficient in the case of diverse and conflicting data. The purpose of this project is to conduct a robust statistical analysis in order to evaluate current operations and to provide such a basis for expansion. In order to assist ACO in determining which factors contribute to the success of their locations, we intend to design a ten-factor model that predicts the success of a new store.

Annual gross profit (AGP) has been selected as the metric to best represent a “successful store.” This metric avoids single-time emergency costs that may have been incurred throughout the year, allowing for the highest degree of standardization across stores. It is worth noting that nearly half of the AGP for the fiscal year 2006 comes from 18 stores, less than 30% of the total number of stores. Referring to Figure 1, it is evident that a large portion of ACO’s annual profit is from a rather small subset of stores.

The model considers the geographic data that ACO has provided, in order to build a model that considers only factors that are directly under the control of ACO at the time of building a new store. Hence, traffic and competition data are the main factors that are included in the model. Marketing data is not factored in, to allow such data to be considered independently of the geographic data.



**Figure 1.** Percentage of Total Profits for Fiscal Year 2006 Group by Profit Rank.

## Multiple Linear Regression

Consider a hypothetical car dealer. A car dealer may tabulate data for each car sold on his lot. Every car has a unique combination of factors; that is, objective data ranging from model year, mileage, and features not available on a standard model; to subjective data, such as customer ratings for “sex appeal” or “fun to drive.” Once this data has been tabulated for all cars sold, the dealer may wish to determine what factors contribute most to the selling price of the car.

The multiple linear regression method is used to analyze the data in order to learn more about the relationship between several independent, or predictor, variables and several dependent, or criterion, variables. In the case of ACO Hardware, as in the case of the car dealer above, the application of the method is limited to a single dependent variable.

Once the parameter values are found, the regression equation is determined. This equation is the best prediction of the criterion variable in terms of the predictor variables available. A full mathematical explanation of the method is available in the Appendix.

## Data Processing

The theoretical underpinning of the regression technique is concerned with idealized data ready to be analyzed. However, raw data often suffers from incompleteness or

codependency (autocorrelation), and must be processed prior to analysis. Without such processing, no statistically meaningful results may be inferred. Clearly, a regression may not be performed on incomplete data, and while regression may be performed on autocorrelated data, the parameters generated may be incorrect, rendering a logical interpretation of the equation impossible.

The raw data supplied by ACO Hardware is not exempt from these problems. Many factors in the raw data were in fact given as formulas based upon other factors, resulting in a much higher degree of autocorrelation than is acceptable. Prior to implementing the regression, these factors were removed. Further investigation of the raw data indicated that there was only partial fiscal year data for three stores. Those three stores were also removed.

Finally, many of the factors were in the form of non-numerical polychotomous data: there are a finite number of discrete states for each factor, where each state is given as string data. The leftmost table in Table 1 shows an example of finite states given as text: the factor “visibility from road” can carry the values “good,” “fair,” or “poor.” Prior to the implementation of a regression, the factor must be transformed into a set of binary variables with only two possible states: “yes” or “no.” In order to use this data, these states must be converted into their respective standard Boolean values, 0 or 1. The process is illustrated below:

**Table 1.** Conversion of non-numerical polychotomous data to a set of binary values.

Visibility from road		Visibility “Good”	Visibility “Fair”	Visibility “Poor”
Good	Conversion →	1	0	0
Fair		0	1	0
Good		1	0	0
Good		1	0	0
Poor		0	0	1
Fair		0	1	0
Poor		0	0	1
Fair		0	1	0
Good		1	0	0
Fair		0	1	0
Good		1	0	0
Poor		0	0	1

Once the data has been properly formatted, the regression is performed with MINITAB, a statistical analysis software package used throughout industry. MINITAB automates the process of determining parameter values and residual values.

## Data and Residual Analysis

The equation output by MINITAB is shown in Table 2. It shows annual gross profit as a combination of metric data and binary data. A more detailed analysis of the variables involved in the regression equation follows in Table 3. Note that the variable names are shortened in Table 3 for the sake of readability.

**Table 2.** Regression Equation Output by MINITAB.

The regression equation is	
<b>Gross Profit</b>	= - 29042
	+ 49.9 Square footage of store
	+ 37.1 \$ repair spent per household
	+ 0.346 Sum of traffic
	+ 89024 Trunk line (y/n)
	- 119389 Grocery Store (y/n)
	- 71148 Drug store (y/n)
	+ 14576 ACE within 4 mile radius (y/n)
	- 69987 Home Depot within 4 mile radius (y/n)
	+ 69552 Road visibility rated "Good"
	+ 89849 Road visibility rated "Fair"
	+ 66514 Directives rated "Good"
	+ 24826 Directives rated "Fair"

**Table 3.** Detailed Variable Analysis Output by MINITAB.

Predictor	Coef	SE Coef	T	P
Constant	-29042	141549	-0.21	0.838
Sq footage	49.878	6.054	8.24	0.000
\$ repair	37.12	86.24	0.43	0.669
Traffic sum	0.3458	0.5004	0.69	0.493
Trunk line (y/n)	89024	36267	2.45	0.017
Grocery (y/n)	-119389	33575	-3.56	0.001
Drug (y/n)	-71148	34445	-2.07	0.044
ACE 4mi radius (y/n)	14576	33332	0.44	0.664
HD 4mi radius (y/n)	-69987	34017	-2.06	0.045
Good visibility	69552	44511	1.56	0.124
Fair visibility	89849	49412	1.82	0.075
Directives_good	66514	44817	1.48	0.144
Directives_fair	24826	40244	0.62	0.540

**S = 126565    R-Sq = 72.5%    R-Sq(adj) = 66.2%**

Some comments must be made on the equation. While the first eight factors listed in Table 3 vary independently, the visibility factors and the directives factor are linked. Since there are three possible states for each factor, only two, "good" and "fair," are

analyzed by MINITAB. Both are considered an improvement over the “poor” state not given in equation. Also, if a store is found to be in one state, it cannot possibly be in another. A more detailed list of the ten factors are given in Table 4.

**Table 4.** List of Ten Factors.

<b>Factor Name</b>	<b>Possible States</b>
Square footage of store	Continuous
Average \$ spent in repair	Continuous
Sum of passing traffic	Continuous
Store lies on a “trunk line”	Yes / No
Grocery store within shopping center	Yes / No
Drug store within shopping center	Yes / No
There is a Home Depot within 4 miles	Yes / No
There is an ACE Hardware within 4 miles	Yes / No
Visibility from road	Good / Fair / Poor
Ability to follow directives	Good / Fair / Poor

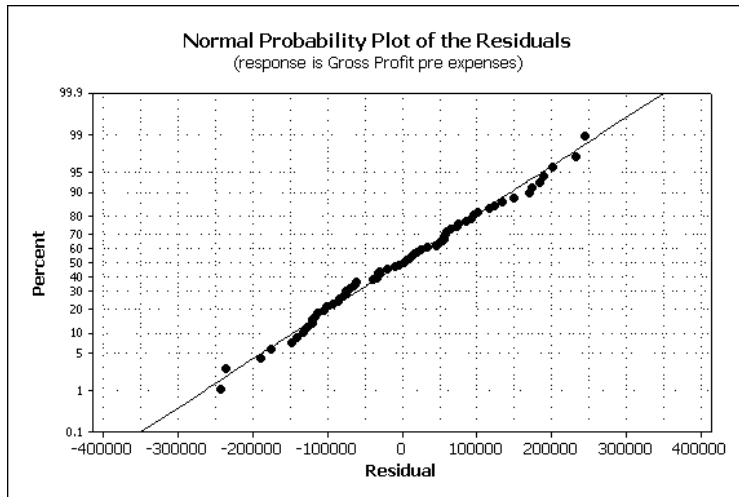
Before beginning a more detailed examination of the ten factors, it remains to examine the strength of the regression. This is done through residual analysis. The smaller the variability of the residual values around the regression line relative to the overall variability, the better the prediction. For example, if there is no relationship between the  $Z$  criterion variable (profit) and the set of  $X_i$  predictor variables (square footage, etc.) then the ratio of the residual variability of the  $Z$  variable to the original variance is equal to 1. If  $X$  and  $Z$  are perfectly related then there is no residual variance and the ratio of variance would be 0. In real-world cases such as this, the ratio falls somewhere between these extremes; that is, between 0 and 1.

From the residual variance the correlation coefficient, commonly known as the  $R^2$  value, is computed. This value may be interpreted in the following manner: if the  $R^2$  value of the equation is 0.4, then the variability of the  $Z$  values around the regression line is  $(1.0 - 0.4)$  times the original variance. In other words, 40% of the original variability in the data is determined by the regression, with 60% residual variability remaining. Thus, the  $R^2$  value is an indicator of how well the model fits the data (e.g., an  $R^2$  value close to 1 indicates that almost all of the variability in the data is determined by with the variables specified in the model).

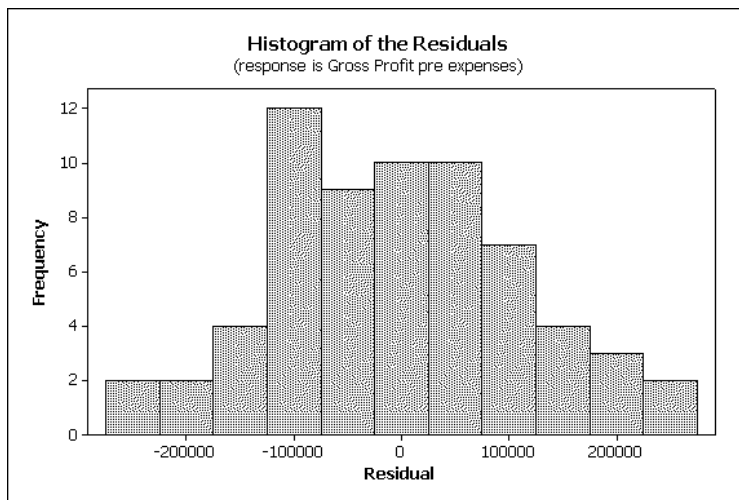
The correlation coefficient for the model developed is lower than an optimal coefficient of 0.8, the original target of this project. Other models have been developed with nominally higher  $R^2$ , but all lack a common-sense interpretation. This model was selected over others due to its logical interpretation, even though the  $R^2$  value is lower than desired.

However, a high correlation coefficient is not the sole indicator of an acceptable regression model. Further analysis of the residual terms must be performed before

accepting a regression model. MINITAB outputs a series of graphs, shown below, that allow for visual inspection of the residual terms, as seen in Figures 2, 3, and 4. The residual terms must be distributed normally, and must not be autocorrelated. Figures 2 and 3 will deal with the normal distribution assumption, and Figure 4 will deal with the autocorrelation assumption.

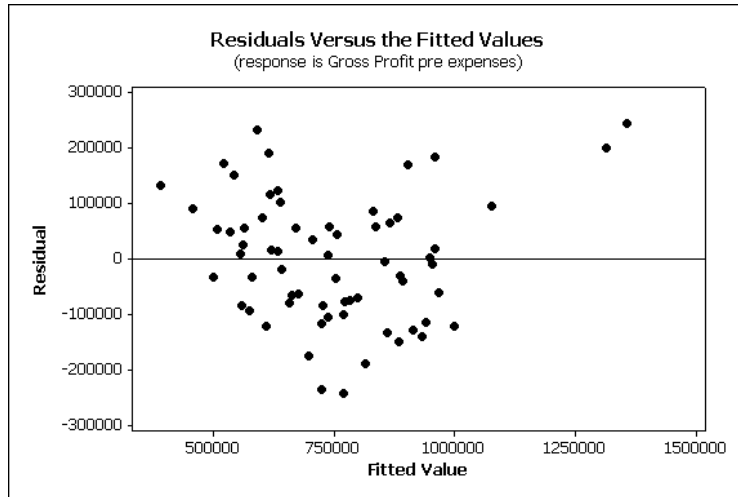


**Figure 2.** Normal Probability Plot. The ideal fit is directly on the line.



**Figure 3.** Histogram of the Residuals. The ideal histogram is normally distributed – a bell curve.

From Figures 2 and 3 it can be seen that the error terms are very nearly normally distributed, well within the acceptable range. Finally, an investigation of whether the residual terms may be written as a function of  $x$  determines whether the error terms are autocorrelated. If the terms are indeed autocorrelated, a discernible “cone” shape will develop as  $x$  tends to infinity. Since Figure 4 does not display this “cone” shape, the error terms are distributed randomly.



**Figure 4.** Residuals versus Fitted Values. The ideal distribution is completely random.

There is an additional assumption in the regression method that is of no concern in this case, the assumption that the error terms do not vary with time. Since all of the data considered is taken as a “snapshot,” there is no time variable the data could possibly vary to. Thus this component of the analysis may be ignored without effect. Hence the residual analysis of the regression model for ACO Hardware is complete, and interpretation may commence.

## Discussion of the Factors

Referring to Table 4, there are three types of factors; those that have continuous states, those that only have “Yes / No” states, and those that have “Good / Fair / Poor” states. Each group will be treated separately. Table 5 shows the first grouping, the continuous factors.

**Table 5.** Factors given continuously.

Factor Name	An additional	Increase in AGP
Square footage of store	Square foot of retail space	\$ 49.90
Average \$ spent in repair	Average dollar spent per year on household repairs and maintenance	\$ 37.10
Sum of passing traffic	Car passing in front of store	\$ 0.346

The model allows for some inferences based upon the coefficients determined by MINITAB. For example, it may be inferred that additional square footage will result in larger profits. Specifically, an additional square foot of space will lead to nearly \$50 in additional profit, according to the model.

Considering traffic, the model determines an increase gross profit by 34.6 cents for every car that passes in front of an ACO Hardware store. Note that these cars are potential customers, not actual customers. The visibility of the ACO brand is enhanced

by an increase in passing traffic, leading to higher annual profit. Since an average of 45,000 cars pass in front of a store, the average increase in profit is nearly sixteen thousand dollars.

More important to visibility of the ACO brand is whether a store is located on a trunk line, a main route of heavy passage. The model indicates that the additional exposure that the “trunk line” designation generates is worth \$89,000. The coefficients for all the “Yes / No” factors are detailed in Table 6 below.

**Table 6.** Yes / No Factors. The change in AGP is given if the factor carries a “Yes” value.

<b>Factor Name</b>	<b>Change in AGP</b>
Store lies on a “trunk line”	\$ 89,024
There is a grocery store within shopping center	\$ (119,389)
There is a drug store within shopping center	\$ (71,148)
There is a Home Depot within 4 miles	\$ (69,987)
There is an ACE Hardware within 4 miles	\$ 14,576

The largest coefficients come from competition factors, heavily determining annual gross profit in the model. The largest factors are “grocery store” and “drug store,” each of them causing ACO Hardware to lose a substantial amount of profit. A detailed interpretation as to why ACO loses as much as it does is beyond the scope of this paper, it suffices to note that the relationship exists, and that ACO Hardware would do best to avoid grocery stores and drug stores when considering future expansion.

The two hardware stores that have the most branches within a close proximity to ACO are Home Depot and ACE Hardware. The model clearly shows the negative affect that proximity to a Home Depot entails. However, it is interesting to note that having an ACO Hardware store within a four-mile radius of an ACE Hardware actually causes an increase in profit in our model. Again, a detailed explanation is beyond the scope of this paper.

The visibility factor presents somewhat of a paradox. Below, Table 7 shows that stores with a visibility rated “good” actually have a lower increase in annual gross profit than store with merely a “fair” rating. This may be explained by the imbalance between the number of stores with each rating. For every store with visibility rated “good,” there are *two* stores rated “fair,” leading to an imbalance in that data. It is this imbalance that is most likely causing the values shown below.

**Table 7.** Visibility Rating.

<b>Visibility rating</b>	<b>Change in AGP</b>
Good	\$ 69,952
Fair	\$ 89,849
Poor	\$ 0

The final factor concerns the central planning ability of ACO Hardware. It is shown in Table 8 that the better a store follows directives from the central office, the better the store does. This information is useful both to the ACO corporate office as well as to each store, as it shows that the corporate office has a planning ability that individual stores would do well to heed.

**Table 8.** Directives Rating.

<b>Ability to follow directives</b>	<b>Change in AGP</b>
Good	\$ 66,514
Fair	\$ 24,826
Poor	\$ 0

## **Conclusion**

A robust statistical analysis of the data collected by ACO Hardware has been performed. After preparing the data, a multiple linear regression was performed, yielding ten factors that influence the expected annual gross profit of a single ACO Hardware location. Though the regression performed resulted in a correlation coefficient of 0.725, less than the optimal 0.8 set as a target, further analysis of the equation shows that the residual terms are well within the acceptable range. Thus the equation developed through the multiple linear regression technique provides ACO a basis from which to plan future expansion.

## **Future Work**

The focus of this project is geographic location, but there is a wealth of marketing data that may be analyzed so that ACO may better know its customers. A model that analyzes this data has already been proposed as Phase II of the ACO Hardware project for MTH844, Projects in Industrial Mathematics, Spring 2007, at Michigan State University.

Also, a different model, using different data, may be built specifically highlighting some of the factors that have little representation in our model, such as “free standing” stores or “rural” stores.

## **Acknowledgements**

The project team would like to thank Mr. John Williamson, ACO Hardware Technology, Mr. Richard Hensch, MSU Industrial Mathematics, and Dr. Peiru Wu, MSU Industrial Mathematics for the opportunity to tackle a significant business problem.

## References

- [1] Long, J. Scott. Regression models for categorical and limited dependent variables. Thousand Oaks, California: SAGE Publications, 1997.
- [2] MacCluer, Charles R. Industrial Mathematics: Modeling in Industry, Science, and Government. New Jersey: Prentice Hall, 2000.
- [3] StatSoft, Inc. "Multiple Regression." Viewed On-line, September 2006:  
<http://www.statsoft.com/textbook/stmulreg.html>.
- [4] Srivastava, Sen. Regression Analysis: Theory, Methods, and Applications. New York: Springer-Verlag, 1990.

## Appendix (Multiple Linear Regression)

In order to perform the linear regression, assume an ideal model where each criterion variable  $Z_k$  may be written as a linear combination of the predictor variables  $X_{k_i}$ . In the case of ACO Hardware, the single criterion variable  $Z_k$  is profit, while the  $X_{k_i}$ 's are factors such as square footage per store, sum of traffic in front of store, etc. Now, allow each criterion variable to be expressed in the form

$$Z_k = \sum_{i=1}^n \alpha_i X_{k_i} + \alpha_0 + \varepsilon, \quad k = 1, 2, \dots, n$$

where the  $\alpha_i$ 's are parameters – the values that correspond to each predictor variable. The equation can then be judged as a combination of predictor variables, each with its own importance.

The values are to be determined by the method of least squares in order to minimize the sum of the squares of the residuals

$$S_r = \sum_{k=1}^m \left[ \left( \sum_{i=1}^n \alpha_i X_{k_i} \right) + \alpha_0 - Z_k \right]^2.$$

Minimize this residual sum by solving for the minimum of the quadratic function of  $n + 1$  variables  $\alpha_1, \alpha_2, \dots, \alpha_n, \alpha_0$ . Taking the partial derivatives of  $S_r$  with respect to each  $\alpha_i$  and setting each to zero, it follows that yields the system of  $n + 1$  linear equations given in terms of the parameter values  $\alpha_1, \alpha_2, \dots, \alpha_n, \alpha_0$

$$\sum_{k=1}^m X_{k_j} \sum_{i=1}^n X_{k_i} \alpha_i + \sum_{k=1}^m X_{k_j} \alpha_0 = \sum_{k=1}^m X_{k_j} Z_k \quad j = 1, 2, \dots, n$$

$$\sum_{k=1}^m \sum_{i=1}^n X_{k_i} \alpha_i + \sum_{k=1}^m \alpha_0 = \sum_{k=1}^m Z_k.$$

In matrix form, the problem becomes solving  $X\alpha = z$ , where  $X$  is the factor matrix,  $\alpha$  is the parameter vector, and  $z$  is the target variable; that is,

$$X = (x_{ij}) = \left( \sum_{k=1}^m X_{k_i} X_{k_j} \right), \quad \alpha = \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_m \end{pmatrix}, \quad \text{and } z = \left( \sum_{k=1}^m X_{k_i} Z_k \right).$$

Once this matrix is obtained, there are a number of ways to compute the solution.

