USING STATISTICAL MODELS FOR MARKET SELECTION

by

Kenneth M. Johnson

Managers responsible for corporate development in growing firms are often called upon to identify promising new markets. Selecting such markets is a complex problem requiring the simultaneous consideration of many demographic, economic, business and competitive factors (Craig, et al, 1984). The need to examine a large pool of possible markets in order to identify those with the greatest potential further complicates the selection process. This article provides an overview of a statistical technique known as general linear modeling and explains how it can be successfully applied to the task of selecting markets.

Using Models to Select Markets

General linear models are particularly appropriate to problems of market and site selection because they offer researchers several important pieces of information. First, they produce a specific revenue prediction for each market of interest. These revenue predictions can then be used to rank the markets. Second, such models generally provide more accurate revenue estimates than can be obtained from simple models because they simultaneously consider the impact of numerous factors on revenue. Third, the contribution of each predictor variable to revenue can be estimated. As a result, the model may be made more efficient by excluding variables which do not improve the prediction of revenue, thereby reducing the expense and time required for data collection. Finally, the statistical technique provides an explicit measure of the quality of the overall prediction model. This provides managers with a means for assessing the accuracy of the model.

General linear models use information about existing markets to produce a statistical model to predict for similar potential markets. Thus, the procedure discussed here is most appropriate for established regional or national firms seeking to expand. It is not appropriate for new firms or those with only a few locations, because they do not have information about enough existing locations to provide a database for modeling. For such firms the checklist or analog approach may be more appropriate (Craig, et. al., 1984). Although not discussed here, the general linear models can also be used for site selection (Lord and Lynds, 1981; Olsen and Lord, 1979). The statistical basis of the general linear model is complex and certainly beyond the scope of this article, however, the following example provides an idea of how the technique is applied to the problem of market selection.

Selecting Markets for Welcome Motels

Consider the problem facing the market development manager of the fictitious Welcome Motel chain. Senior management is planning a major corporate expansion and she has been asked to identify sites with high potential to support new Welcome units. The firm is already well established nationally. Senior management wants to increase Welcome's presence in middle markets, defined as those with populations of 50,000 to 250,000. Such markets are either Metropolitan Statistical Areas (MSA's) or, in nonmetropolitan areas, a county of at least 50,000 containing a city of 25,000. Welcome already operates units in 100 such middle markets. Some of these units are doing quite well, others not so well. The market development manager must decide which of the numerous middle markets not yet served offer the greatest potential for the corporation.

The manager has or can obtain data on the performance of the 100 Welcome units already operating in middle markets. When such information is combined with demographic, economic and business information for these markets, the resulting database can be examined using general linear modeling. If the resulting model provides accurate estimates of performance in the existing locations, it can then be used to predict performance in the potential markets.

Phase I: Modeling Existing Locations

A project such as this for Welcome should be executed in two phases. Such a phased approach provides senior corporate staff with the opportunity to monitor the progress of the project and terminate it if the model does not prove to be satisfactory. A number of steps within each phase will be discussed, but it is important not to lose sight of the overall goal -- which is to generate revenue estimates for potential markets in order to identify those with the greatest promise. In this first phase of the project, data about existing markets is used to develop a statistical model which may then be used to estimate revenue for potential locations.

The first step in the process is to select a set of existing locations similar to the potential markets. For Welcome, the 100 existing middle market locations are the most likely candidates. These existing markets should be screened to exclude any very unusual markets. However, the existing markets used should include both those that are doing well and those that are doing poorly. It would be a serious error to exclude underperformers because the model must be sensitive to what exerts a negative impact on performance as well as what exerts a positive influence. Major markets (those with populations greater than 250,000) should not be included in the modeling process because the dynamics in such markets are quite different from those in middle markets, particularly in the accommodations industry.

The next step is to identify potential predictor variables. This is a critical step because if important predictor variables are excluded, the model will not make accurate revenue estimates, perhaps jeopardizing the entire project. The variables finally identified reflect several major factors believed to influence

the revenue potential of a Welcome location (Figure 1). The first is the character of the local population and economy. Welcome staff know much of their business comes from the local community whether through accommodations for visitors, weddings, business meetings or whatever. A second key factor is the size of transient population. A substantial proportion of motel revenue comes from individuals in the area temporarily for business, vacations, special events, and so forth. The extent and quality of competition from other motels and hotels represents a third important influence. Welcome locations without extensive competition can certainly be expected to do better than Welcome units in similar locations with significant quality competition. A final factor is the quality of the management at the Welcome location. A well managed Welcome unit may do much better than a unit in a similar market with lackluster management. In all, sixty-one predictor variables were identified which together measure these four factors. Relatively few of these variables will remain in the final model, but it is best to start with a large pool of variables and eliminate the ones that do not prove useful based on statistical analysis. At this point the greatest concern is not to exclude a potentially useful variable.

Once Welcome collects and prepares the database on its 100 existing middle market locations, statistical analysis can begin. Such analysis focuses on identifying a subset of the 61 predictor variables which estimate the revenue for the existing locations with reasonable accuracy. Although statistically complex, the modeling procedure is conceptually simple. What is needed is a statistical abstraction of the local market which responds to differences in the factors which influence revenue. The modeling procedure provides such an abstraction by using a weighted combination of predictor variables to account for revenue differences among the existing markets.

In the case of Welcome Motels, eight of the original 61 variables were particularly important in predicting revenue for existing locations. Among these, population and number of firms employing 100+ workers reflects the character of the local area (see Table 2 following). Hotel sales also partially reflect the character of the local market, but is primarily a measure of the transient population. The size of the transient population is reflected in the number of daily nonstop commercial flights, number of vehicles entering the county, and the number of major events annually (i.e. festivals, conventions). The impact of competition is reflected in the number of major competitors Welcome faces in each market. Finally, the quality of management at the Welcome unit in each market is needed to make accurate revenue predictions.

The decision to use these eight variables to estimate revenue represents a trade off between accuracy and efficiency. Including additional variables marginally improves the predictive power of the model, but it makes the model less efficient and more sensitive to factors unique to one or a few markets which may not be duplicated among the potential sites. What is required here are the factors that are general and therefore

Figure 1

Factors Expected to Influence Revenue Potential

likely to influence the revenue potential of a Welcome unit in any city.

The weight assigned to each of the eight variables for predicting revenue is determined by the statistical model. In principle, each weight reflects the difference in revenue expected between two locations that differ by one unit on the specific predictor variable if the two locations are in the same position on all other predictors. For example, if two existing Welcome locations are the same on all the predictors but flights and city A has one more daily nonstop than city B, then the Welcome location in city A is expected to generate \$140,300 more revenue annually. This formal interpretation of regression weights is of limited utility in applied settings because if cities differ on one important predictor, they typically differ on others as well. Although this complicates the analytical task of deriving the appropriate statistical model, it does not influence the actual prediction of revenue for each market.

Evaluating the Quality of the Model

For any given Welcome location there is likely to be a difference between the actual revenue generated and the revenue predicted for that location by the model. Such variation is expected given the multiplicity of variables which influence how well an individual location does. The modeling procedures are designed to minimize the differences between the predicted and actual revenue for all locations. Once the best model for predicting revenues for existing markets has been selected, the quality of the predictions derived from it must be evaluated. The key point is whether revenue predictions derived from the model are accurate enough to aid the firm in planning its expansion. As noted above, there will be differences between the predicted and actual revenue for the existing locations used to derive the model. The managers and research team must determine whether the model predicts revenue for existing locations accurately enough to justify its use in selecting new markets.

The quality of the predictions generated by the model can be assessed with a number of formal statistical measures. The most common of these is the squared multiple correlation coefficient (R-square). Here it reports the proportion of all the variation in revenue among the locations accounted for by the prediction model. R-square varies from zero to one with values closer to one suggesting more accurate solutions. In the case of Welcome, R-square is .86 suggesting the model provides quite accurate predictions of revenue for the existing locations.

The abstractness of statistical measures such as R-square limit their utility in evaluating the quality of models in applied settings. A number of other techniques may be more appropriate in such settings. One can, for example, calculate the percentage difference between the actual and predicted revenue for each market and then average them over all markets. In the case of Welcome, the average deviation of the predicted from the actual score was 20 percent. Thus, a location with a predicted revenue of \$1,000,000 per year could reasonably be

expected to have sales between \$800,000 and \$1,200,000. An even more useful technique is available in Welcome's case because the market development director can specify revenue thresholds for what constitutes unsuccessful (revenue less than \$1,000,000), moderately successful (revenue of \$1,000,000 to \$2,000,000), and highly successful (revenue of \$2,000,000 plus) locations. Using such information, a crosstabulation is constructed comparing the predicted and actual revenue of the 100 existing locations (Table 1).

Table 1

Comparisons of Predicted and Actual Revenues
for Current Welcome Locations in Middle Markets

		Predicted 1983 Revenue		
		Inadequate Revenue	Moderate Revenue	High Revenue
Actual 1983 Revenue	Inadequate Revenue	80.0%	32.7%	3.0%
	Moderate Revenue	20.0	53.0	18.2
	High Revenue	0.0	14.3	78.8
		<u>100.0%</u> (20)	<u>100.0%</u> (37)	<u>100.0%</u> (43)

Where: Inadequate Revenue = Less than \$1,000,000
Moderate Revenue = \$1,000,000 to \$2,000,000
High Revenue = Greater than \$2,000,000

This crosstabulation clearly illustrates the ability of the model to discriminate between unsuccessful and successful locations. For example, in 80 percent of the cases where the model predicted inadequate revenues for successful operation, the actual revenues generated were, in fact, inadequate. In contrast, nearly 80 percent of the locations identified as highly successful by the model were, in fact, highly successful and only 3 percent were unsuccessful. A presentation stressing such applied criteria is often more salient to managers evaluating the quality of the model than is one focused on the more abstract statistical measures.

Evaluation of the model at the aggregate level should be supplemented by a case by case comparison of the predicted and actual revenues. Such review serves the dual purpose of identifying extreme outliers (cases for which the predicted and actual revenue are quite different) and giving managers a better feel for the results of the modeling procedure. Often the outliers are well known to the managers and the fact that they

appear as outliers may actually increase management's confidence in the model. Occasionally such a review will uncover a new factor that accounts for such outliers, in which case it can be incorporated into the model thereby improving the quality of the solution. In any event, a case by case review of the locations is an important way of gaining an indepth understanding of the strengths and limitations of the modeling procedure.

Phase II: Predicting Revenues for Potential Locations

After reviewing the results of the model for existing locations, a decision must be made regarding whether the project should be continued. In the case of Welcome, management was satisfied that the model produced reasonable predictions of revenue and that such predictions will be useful in selecting new locations.

Phase two begins with the identification of a pool of markets that can be examined as potential locations. Care must be exercised in selecting possible markets because those selected must have characteristics reasonably close to the existing markets used for modeling. Thus, because the existing Welcome locations used to develop the model were all middle size markets (50,000 - 250,000), those included in the pool of potential locations should be of roughly the same size. Review of available demographic data identified 150 markets of the proper size to be examined as possible Welcome markets.

Having selected the pool of possible markets, data must be collected for each. Note that it is only necessary to collect data for the variables to be included in the modeling procedure. This dramatically reduces the time and cost of collecting data. In the case of Welcome, information need only be collected on the 8 predictor variables used in the model.

A problem certain to arise in data collection is that of missing data. In some markets data on one or more of the predictor variables will not be available either because some agencies do not report the data required or because of Census Bureau suppression to avoid violation of confidentiality. Often a reasonable estimate of such missing values can be made allowing the case to be included. A market should only be eliminated from the pool if data is missing on several variables.

Once the database is constructed for the potential locations, the model developed for existing markets can be used to generate revenue estimates for each of the potential markets. Consider the example of Eau Claire metropolitan area in Wisconsin, it was identified as a potential Welcome site because it had an area population in excess of 50,000 and a city (Eau Claire) of more than 25,000. Having collected data on each of the relevant variables we now insert the appropriate values into the model and produce a revenue prediction for the potential market (Table 2). The value for each predictor is multiplied by the weight assigned to it to determine what contribution it will make to the model. For example, the population of 130,507 is multiplied by the weight of 4.1 to provide an estimated contribution to revenue of \$535,079. In reviewing the relative contribution of each

TABLE 2: EXAMPLE OF PREDICTION MODEL
WELCOME HOTELS
EAU CLAIRE, WISCONSIN

REGRESSION

VARIABLE	VALUE	WEIGHT	CONTRIBUTION
POPULATION 1980	130507	4.1	\$535079
HOTEL SALES (000'S)	6876	30.8	211780
# FIRMS EMPLOYING 100+	48	5147.0	247056
# MAJOR EVENTS	8	11457.0	91656
# OF NONSTOP FLIGHTS	3	140300.0	420900
# VEHICLES ENTERING AREA	36000	2.7	97200
# MAJOR COMPETITORS	4	-97843.0	-391372
QUALITY OF MANAGEMENT	3 (GOOD)	43354.0	130062
CONSTANT	1	101453.0	101543
			\$1443904

CHARACTERISTICS OF THE MODEL
STANDARD ERROR OF ESTIMATES = 328000
SQUARED MULTIPLE REGRESSION COEFFICIENT = .86
SAMPLE SIZE = 100

Where:

POPULATION 1980 = Area population in 1980 (Census of Population)

HOTEL SALES = Total Hotel and Motel Revenue in '000s (Census of Service Industries)

FIRMS EMPLOYING 100+ = # of firms employing more than 100 workers (County Business Patterns)

MAJOR EVENTS = # of festivals, conventions and other tourist events (Chamber of Commerce)

NONSTOP FLIGHTS = # of daily nonstop flights by commercial carriers (Official Airline Guide)

VEHICLES ENTERING COUNTY = # of vehicles entering area daily (State Department of Transportation)

MAJOR COMPETITORS = # of motels with 100+ rooms and toll free reservations (Tourist Industry Sources)

QUALITY OF MANAGEMENT = Rating of local Welcome management from 0 (low) to 4 (excellent) (Welcome Operations Staff)

CONSTANT = Coefficient derived from statistical modeling and set to 1 for each location

variable, one might note the particularly strong influence of airline flights. Each additional flight can be expected to contribute \$140,000 dollars to annual revenue other things remaining equal. The sensitivity of the model to airline flights undoubtedly reflects the strong relationship between air travel and the number of transients coming to the area. Note also that an estimate of the quality of management has been included in the model even though no location yet exists in Eau Claire. This predictor allows Welcome executives to better understand the impact of management quality on revenue.

The predicted revenue for each of the other 149 potential markets would be calculated in a manner similar to that for Eau Claire. The markets can then be ranked based on predicted revenue and the most promising among them selected for further detailed review by the market development staff. In this regard, the use of rankings is valuable in that it tends to mitigate the fact that the specific revenue estimate for each potential market is likely to contain some error. Thus, while it is unwise to expect a given potential market to generate exactly the amount of revenue predicted by the model, it is likely its rank relative to other potential markets is quite accurate.

Having generated predictions and ranked the potential markets, the results should now be reviewed both by the senior managers responsible for strategic planning and by the field staff charged with making the detailed investigations of the most promising markets. In the course of such meetings some revision of the rankings is likely to occur as those with first hand knowledge of the markets review the predictions. The field staff, in particular, are often sensitive to nuances of the local markets which are not evident in the statistical abstractions of the modeling process.

Among the 150 potential markets 40 had predicted revenues high enough to warrant further investigation. The Market Development manager assigned these markets to her field staff for detailed review which led to a decision to purchase sites and construct motels in 33 markets. If the units perform at the level predicted by the model, they will generate over 80 million dollars in additional revenue each year for the corporation.

The success of the model in predicting revenue for existing middle market locations also encouraged Welcome managers to use the techniques to evaluate the performance of existing locations. In this regard, the model can be used to identify under-performing markets so that corrective action can be taken sooner. In a similar fashion, markets which are performing above expectation can be examined to determine what caused them to be successful so that the techniques can be applied systemwide. Management plans to introduce such model based evaluation procedures among both its middle and major market locations.

Limitations

Statistical models of the type discussed in this article are powerful aids to a corporation intent on expanding into new markets, but they are not without limitations. First, the models

are only as good as the decisions made in constructing them. If important predictor variables are not included, the model may be inaccurate. A more subtle problem is that of the similarity between the existing markets used to construct the model and the potential markets for which predictions will be made. If there are significant differences between these two groups, the model may not predict revenues properly for the potential markets.

Second, the construction of the statistical models is itself an extremely complex task. Discussion of the technical issues related to model construction have purposely been avoided here assuming either that the reader is well enough versed in these techniques to immediately see their application, or will seek the help of those who are. The statistical subroutines for general linear models are readily available in most mainframe statistical packages (SPSS, SAS), and in the newest generation of packages for micros (SPSS-PC, SYSTAT). However, there is more to such model construction than simply running the data through the program. Many of the decisions required in developing models such as those considered here require a great deal of judgement complemented by a detailed knowledge of statistics and the data available. An inexperienced researcher running the data through the default options of a stepwise regression procedure uncritically could well produce a flawed model that will eventually cost the corporation dearly.

A third limitation is that predictions derived from the model must be interpreted with caution. Even when the overall performance of the model is quite good, the actual revenues a location will generate will vary from what the model predicts. Generally, the difference between predicted and actual revenues will be modest, but in a few markets the differences may well be substantial. The major point here is that it is inappropriate to put heavy stress on the actual dollar revenue predicted for a potential market. Of much greater concern is the position of the market relative to other potential markets.

Finally, statistical models do not negate the need for careful work by managers and the field staff. The modeling procedures provide a means to estimate the revenue potential of a large number of markets and identify the most promising among them. However, the final decisions about which markets to enter must rest on assessments by managers in consultation with the field staff.

Conclusion

The modeling procedures outlined are designed to facilitate expansion by regional and national firms by allowing them to use information about their existing markets to identify promising new markets. Using multiple predictor variables, the models are able to more closely approximate the complexity of the actual market environment thereby allowing for more accurate predictions. Such models provide managers with the means to efficiently examine large numbers of potential markets and identify those with the greatest promise. Managers are then able to concentrate their limited staff and resources on a detailed examination of

the most promising of the potential markets. Used intelligently in combination with the expertise of the field staff and the judgement of senior corporate managers such models provide a firm with the information necessary to expand into new markets successfully.

Sources

Craig, C. S., A. Ghosh and S. McLafferty. 1984. "Models of the Retail Location Process: A Review." Journal of Retailing, 60(1):5-36.

Johnson, K. M. 1985. "Modeling to Find Markets." American Demographics, 7(11):40-43, 54-55.

Lord, J. and C. Lynds. 1981. "The Use of Regression Models in Store Location Research. A Review and Case Study." Akron Business and Economic Review, 10:13-19.

Olsen, L. and J. Lord. 1979. "Market Area Characteristics and Branch Bank Performance." Journal of Bank Research, 10:102-110.]