ELSEVIER

Int. J. Production Economics 58 (1999) 265–276

international journal of
production
economics

Explaining beer demand: A residual modeling regression approach using statistical process control

Murat Koksalan[a,*], Nesim Erkip[a], Herbert Moskowitz[b]

[a] *Industrial Engineering Department, Middle East Technical University, 06531 Ankara, Turkey*
[b] *Krannert School of Management, Purdue University, West Lafayette, IN 47907, USA*

Received 11 July 1997; accepted 6 August 1998

Abstract

We develop a medium-term model as well as a short-term model for understanding the factors affecting beer demand and for forecasting beer demand in Turkey. As part of this specific model development (as well as regression modeling in general) we propose a procedure based on statistical process control principles (SPC) and techniques to (1) detect nonrandom data points, (2) identify common missing, lurking variables that explain these anomalies, and (3) using indicator variables, integrate these lurking variables into the model. We validate our proposed procedure on several test examples as well as on the medium-term beer demand model. Both the medium and short-term models yield very satisfactory results and are currently being used by the company for which the study was conducted. In addition to the residual modeling regression approach developed using SPC, a major contribution to the success of the project (and the modeling in general) is the mutual collaboration between analyst and client in the modeling process.

Keywords: Regression; Forecasting; Statistical process control

1. Introduction

We present an application study made for a private beer company to understand the factors affecting beer demand and to forecast beer demand in Turkey. In addition to presenting the models and their results, we also discuss the difficulties encountered in this application study.

There are two private beer companies and a state enterprise that brew beer in Turkey. The market share of the state enterprise as well as beer imports are very small. The two private beer companies, on the other hand, synchronize the timing and magnitude of their price increases. Due to the high rate of inflation in the country, the companies increase beer prices several times a year.

The project was initiated by the project development department of the company. Though they were interested in forecasting beer demand, their main motivation in this study was to understand

* Corresponding author: Tel.: + 90 312 210 2287; fax: + 90 312 210 1268; e-mail: koksalan@metu.edu.tr.

0925-5273/99/$ – see front matter
PII: S0925-5273(98)00207-2

the factors affecting beer demand. They also wanted to test the intuition of their sales personnel and educate them toward a more scientific approach for understanding the nature of beer demand. We had strong support and cooperation of the project development department of the company throughout the study. We benefited from their insights in the interpretation of the results and received substantial help in data collection. Our clients had strong beliefs regarding the importance of some of the factors which in most cases turned out to be correct. However, it was not easy to convince them that several of the stated factors were not appropriate because of their nature. It also came as a surprise to them when the model results revealed that some other factors did not have an important effect on beer demand. We further discuss these factors in the next section.

Linear regression models were developed to explain and forecast beer demand. One model tries to explain the yearly per capita beer consumption in terms of various independent variables. Another model attempts to capture the short-term effects by considering the monthly beer consumption as the dependent variable. A statistical process control (SPC)-based procedure is developed and implemented to overcome difficulties related with potential missing, lurking variables. The procedure is useful when there is a poor fit or when the residuals point out possible violations of the regression model assumptions.

In Section 2 we discuss the factors considered, based on a survey conducted among the sales personnel of the company. In Section 3 we present the medium-term model, the encountered difficulties, and the SPC-based approach we use to detect potential missing variables based on the residuals. We discuss the short-term model and its results in Section 4 and present our conclusions in Section 5.

2. Important factors

To identify the factors that may affect beer demand, a survey was conducted among the sales personnel in different regional sales departments and the managers in the headquarters of the company. The survey was conducted by the project development department in cooperation with the project team of this study. Twenty factors were included in the survey and the surveyed individuals were asked to rate the factors from 1 to 10, where a higher score corresponded to a higher degree of importance. They were also allowed to list any additional factors they considered relevant.

The project development department prepared a report explaining the results of the survey. The responses obtained from different regional sales departments and headquarters were mutually consistent. In our study we only excluded the factors that were considered unimportant by all departments with the exception of several important factors as explained below.

Quality of the beer and promotion activities were two of the excluded factors that were considered important by company personnel. These factors were not included in our analysis because there had not been any detectable changes in the beer quality and there were no promotion activities during the investigated period. Another factor that was considered important and was not included in the study was the number of sales points. We argued, for the most part, that demand would trigger opening new sales points rather than sales points creating demand. In other words, whenever there is a significant demand for beer at a region, new sales points would be opened there. New sales points may also create some additional demand, but we thought that this would be relatively small in general. On the other hand, using the number of sales points as an independent variable would falsely appear to explain a significant portion of beer demand. Our clients were convinced, after some discussions, to exclude this factor from the study. Advertisement is another factor not included in our study for the following reasons: first, there was not too much difference in the budget used for the advertisement activities (such as sponsoring sports clubs, cultural activities, direct advertisement via the media, etc.) during the studied period. Secondly, each advertisement activity has a different marginal effect that is difficult to measure. Moreover, it is not clear how long an effect advertisement has on beer demand.

Apart from the above factors, all the remaining factors are included in our study. Some of these

factors are used in the medium-term model and some in the short-term model. Some are used directly as independent variables and some are combined into more meaningful composite independent variables. We give the details of the models in the next sections.

3. Medium-term model and modeling the residuals

With this model, we try to understand the factors affecting beer consumption of an average individual in a year. Initially, our clients were interested in a model that explains the total yearly demand of the country in terms of independent variables. After some discussions we agreed that per capita beer consumption would be a more meaningful dependent variable. Eventually, we ended up using both dependent variables separately, however, we will only discuss the per capita consumption model here.

One difficulty we encountered was the dearth of data points. It would not make too much sense to consider too many years in the past, since it would cover periods that are substantially different from each other in terms of transitions the country has undergone and their effects on the beer consumption habits of people. Then we thought of considering different cities as individual data points. This not only created an abundance of data points, but also enriched the models by widening the range of values of dependent and independent variables. However, in disaggregating the model into individual cities we make additional assumptions. Specifically, we assume that the effect of the independent variables for each city is approximately the same, and that there is no need for additional variables to capture the differences between cities. The violation of these assumptions may effect the validity of the model. We further address these issues in conjunction with the results and the variations of the model.

3.1. Description of the model

Following are the variables used in the model. We present more detail about the variables in the appendix.

Dependent variable:

PCS_{it}: per capita beer sales in city i in year t (over the urban population at or above age 18).

Independent variables:

$PCGDP_{it}$: (per capita gross domestic product in city i in year t) = per capita gross domestic product for the urban part of city i in year t/real price of beer in year t. This variable is a measure of how many liters of beer can be purchased with average income. It combined two factors: the price of beer and the average purchasing power of consumers.

$PCTOUR_{it}$: per capita effect of the tourism factor for city i in year t. This variable considers the countries tourists are coming from together with their lengths of stay and average beer consumption statistics in those countries. It combines information on foreign tourists, domestic tourists and the consumption habits of tourists into a single variable.

$CLIM_{it}$: climate measure of city i in year t. This variable is a measure of average temperature.

$CLIMVAR_{it}$: variation of climate measure in city i in year t. This variable is a measure of monthly deviations in temperatures from long-term averages for that month.

$CLIMNEGV_{it}$: negative variation of climate measure in city i in year t. Measures only negative deviations from long-term monthly averages. (The climate related variables are not found to be correlated, hence can be used in a given model.)

$DVLPNDX_{it}$: development index for city i in year t. Among several development indices, we selected the index that gives social factors a relatively larger weight since economic factors are also considered by other variables.

SB_{it}: a measure of effective distribution of beer from the state enterprise in city i in year t.

CPI_{it}: consumer price index in city i in year t. This variable is included to reflect the purchasing power of consumers in city i in year t.

We consider only the urban population in the model because alcoholic beverages were only allowed to be sold in regions having a municipality (during the period covered in this study). Some of

The Model:

$$PCS_{it} = \beta_0 + \beta_1 PCGDP_{it} + \beta_2 PCTOUR_{it} + \beta_3 CLIM_{it} + \beta_4 CLIMVAR_{it} + \beta_5 CLIMNEGV_{it} + \beta_6 DVLPNDX_{it} + \beta_7 SB_{it} + \beta_8 CPI_{it} + \varepsilon_{it}$$

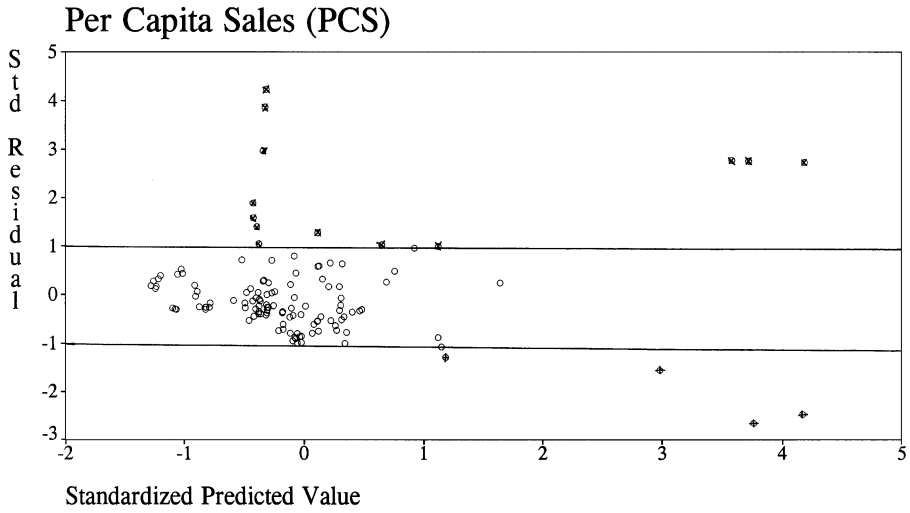

Fig. 1. The residual graph of the medium-term regression model.

the cities have been combined into a single data point, because a more reliable beer sales value could be obtained for the combined cities. We will still refer to these data points as cities.

Several independent variables have been used to represent climate because our clients felt that it has an important effect on beer sales. In addition to the above independent variables, we also considered the ratio of beer price to average hard liquor price and to average soft drink price as independent variables considering hard liquor as well as soft drinks as substitutes for beer. However, the price ratios were the same for all cities (as prices of different beverages do not change between cities) and only varied between years. The corresponding independent variables had only several distinct values and this, in some cases, caused the variables to have significant but counter-intuitive effects. We also thought that, rather than the ratios of yearly average prices, the short-term changes of relative prices would be more likely to affect beer demand. Therefore, we excluded these variables from the current model and used them only with the short-term model as discussed in Section 4.

3.2. Results

The regression model was run for a three-year period from 1989 to 1991, since the sales figures for the cities were most reliable for this period. There were 42 data points (cities) for each year, yielding a total of 126 data points. None of the pairwise correlation estimates between independent variables were very high, eliminating concerns of multicollinearity.

We ran the regression model by the backward elimination of insignificant independent variables. The value of the adjusted R^2 turned out to be 0.60 indicating that about 60% of the variation in the dependent variable can be explained. We present a scatterplot of standardized residuals vs. predicted sales in Fig. 1, which shows an undesirable trend indicating that the variance may not be constant (the horizontal lines at ± 1 standard deviation points and the highlighted data points are referred to in the next section). Our various efforts of transforming the data did not lead to an improvement. Thinking that increasing variance could be caused by missing variables, we developed and

implemented an SPC-based approach discussed below to diagnose the existence of such a possibility.

3.3. Modeling the residuals

Mandel [3] developed a regression control chart in an application in the postal service. A linear regression model was used to explain the man-hours used by the pieces of mail handled. The regression estimates were obtained using data corresponding to time periods having no peculiarities. Then using $\pm 2\sigma$ values, control limits around the regression line were created to be used to detect future points where the process would appear out of control.

In this paper we use a regression control chart type of approach along with indicator variables to detect and incorporate missing, lurking variables into the regression model by eliminating the undesirable trends observed in the residuals, and demonstrate its use both on the beer application considered and on some randomly generated example problems.

3.3.1. The procedure

Consider a regression model where the residuals show a nonrandom pattern, a clustering of points, or an undesirable trend and/or where the adjusted R^2 value is perhaps lower than expected. Could this be due to some unknown, missing variables? Would it be possible to find such variables that, if they existed, would substantially improve the model? If one can find possibly one or two such hypothetical variables, then it would seem worthwhile to search for the true missing variables. The unidentified variables can also pinpoint the data points for which these missing variables take values substantially above or below their mean values. Common characteristics associated with these points help to identify the missing variables that should be included in the model.

In order to do the above, we take the data points that are more than k standard deviations from the corresponding predicted values. In Fig. 1, for example, we draw the control limits for $k = 1$ to show the data points that are outside these limits. We define indicator variables I_1 and I_2 such that I_1 takes on a value of 1 for all data points that are at least k standard deviations larger than the corresponding predicted values and 0 for all other data points. Similarly, I_2 takes on a value of 1 for data points that are at least k standard deviations smaller than the corresponding predicted values. In the example of Fig. 1, the data points for which variables I_1 and I_2 take on values of 1 are marked with '$\times$' and '$+$', respectively, for $k = 1$. If the resulting regression model leads to a substantial improvement, then one may be justified to search for the real variables that are missing from the model.

Example 1. Consider that there exists a true model $Y = 0.8X_1 + 1.0X_2 + 1.2X_3 + \varepsilon$ where ε is a normally distributed random variable with mean 0 and variance σ^2. We randomly generate 20 values for X_1, X_2, X_3 and ε to obtain the data points. We generate X_1, X_2 and X_3 independently from identical normal distributions with mean 20 and variance 25. We generate ε using $\sigma^2 = 25$. Then we run several regressions:

Model 1a. Using all independent variables.

Model 1b. Using X_1 and X_2 and excluding X_3.

Model 1c. Using X_1, X_2, I_1 and I_2, where I_1 and I_2 are indicator variables derived from the residuals of Model 1b with $k = 1$ standard deviations.

The regression results show that the adjusted R^2 is 0.85 with Model 1a whereas it is 0.57 with Model 1b. The graph of residuals of Model 1b also shows a clear undesirable trend. Using the indicator variables in Model 1c, we see that the adjusted R^2 increases to about 0.89. The residual graph of Model 1c shows an improvement over that of Model 1b though it is not as good as that of the true model. A closer look at the data points shows that I_1 takes on a value of 1 mostly when X_3 is above its mean and I_2 takes on a value of 1 mostly when X_3 is below its mean. This indicates that the approach has been successful in capturing the data points where the effect of X_3 cannot be explained by the intercept. This result would indicate that there is a significant potential for improving the results if one or more appropriate missing variables can be identified.

The Model:

$$PCS_{it} = \beta_0 + \beta_1 PCGDP_{it} + \beta_2 PCTOUR_{it} + \beta_3 CLIM_{it} + \beta_4 CLIMVAR_{it} + \beta_5 CLIMNEGV_{it} + \beta_6 DVLPNDX_{it} + \beta_7 SB_{it} + \beta_8 CPI_{it} + \beta_9 VACHOME_{it} + \varepsilon_{it}$$

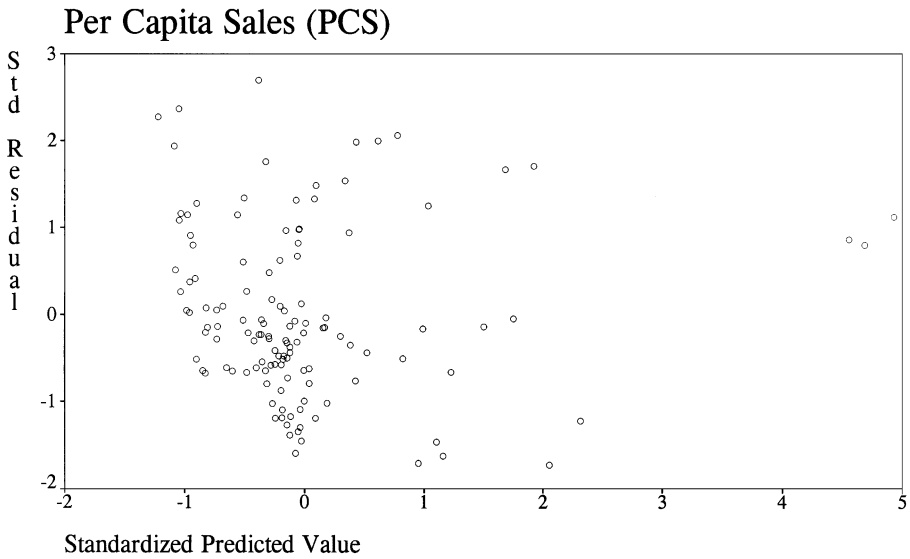

Fig. 2. The residual graph of the model with indicator variables.

Example 2. This example is similar to the previous one except that we use $\sigma^2 = 1$ and a total of 100 data points. Models 2a–c, respectively, correspond to the regression of the true model, the regression when X_3 is excluded, and the regression using X_1, X_2, I_1 and I_2. In this case, the residual plots do not differ too much but the adjusted R^2 could be significantly improved using the indicator variables over the case when only X_1 and X_2 are used. The respective adjusted R^2 values for Models 2a, b, and c are 0.99, 0.52, and 0.86. This again implies that there is substantial room for improvement by finding appropriate missing variables.

3.3.2. The beer demand application

Now, we implement the above procedure for the case of our application. We again use I_1 and I_2 for data points that are at least 1 standard deviation above and below the corresponding predicted values respectively. The results show an improvement in the adjusted R^2 to 0.88 (from the original value of 0.60) as well as an improvement in the pattern of the residual graph.

All of the above analyses merely indicate that there is room for model improvement by the inclusion of one or more appropriate missing variables and the identification of the data points that are most likely affected by them. This motivated the need for our clients and us to carefully examine the data points where the indicator variables took the value of 1 and try to determine what is common to these points. On observing the mentioned data points, we realized that in most cases if the prediction for a city was substantially off in a given year it was also substantially off in other years. We could not find a missing variable that would explain all the data points where indicator variables had a value of 1. However we realized that in four of these cities there was a factor that could not be accounted for. For the independent variable used for tourism, we obtained the information from the publications of the Ministry of Tourism. The information was based on the tourists (domestic and foreign) staying at commercial facilities. In the identified four cities, however, there were many vacation homes where domestic tourists spent their

vacations coming from other cities. We could not find information to create a quantitative variable that would take into account the number of vacation homes. Instead, we approximated this information using an indicator variable (VACHOME) that took a value of 1 for these four cities in each of the three years considered. This regression model yielded an adjusted R^2 value of 0.86. The residual graph of this model is given in Fig. 2. Note that the undesirable residual trends of the original model have disappeared. The independent variables corresponding to the development index, per capita domestic product, tourism, and the vacation homes had a significant positive effect on the beer demand as would be expected. The other significant effect was that of the effective distribution of the state enterprise beer which showed, counter intuitively, a negative effect on beer demand. This variable was originally an indicator variable that took on a value of 1 in a small number of cities, where the beer from the state enterprise was effectively distributed. When we combined cities together, we took the weighted averages of the independent variables, and the variable corresponding to the effective distribution of state enterprise beer took on values other than 0 or 1. Nevertheless, only several data points had nonzero values for this variable and therefore the indicated contribution of this variable to the beer demand by the regression model may not be very reliable.

Of the independent variables considered, the climate-related ones and the consumer price index turned out not to have a significant effect on beer demand. Only the variation in the temperature ($CLIMVAR_{it}$) showed a small negative effect that was significant at the 0.08 level. The climate, in general, is assumed to have a substantial effect on beer demand. Many people, perhaps, can relate from personal experience to drinking more beer when it is hot. Our clients also strongly believed that the climate would have an important effect on beer demand. When the model solution did not show the climate measure ($CLIM_{it}$) to be effective, our clients suggested using a variation type variable. They argued that maybe people do not react to absolute temperature values as much as they react to changes in the temperature values they are accustomed to. This led us to develop the variable $CLIMVAR_{it}$, which turned out to be only slightly significant. Our clients next suggested that negative deviations from average temperatures may have a decreasing impact on beer demand even if positive deviations do not have an increasing impact. The resulting variable, $CLIMNEGV_{it}$, turned out to be insignificant. In the interpretation of the results we argued that although climate may be an important variable, it would be very hard to detect its effect in a model that uses aggregate yearly data. Even though we used deviations in temperature from long-term monthly averages in the variational variables, the data would tend toward the central values when aggregated over the 12 months of a year. We further discuss the effect of climate in conjunction with the short-term model in the next section.

In addition to using an indicator variable for cities having many summer houses, we ran a model incorporating the interaction of the indicator variable with each of the other variables. This allows the independent variables to have different effects (slopes) for the cities where there are summer houses. We summarize the results of the original model, the model having two indicator variables for data points that have large residuals (Variation 1), the model having an indicator variable to represent cities that have summer houses (Variation 2), and the model having interactions (Variation 3) in Table 1. We present the adjusted R^2 values as well as the type of effect each independent variable has in each model. The table shows the substantial improvements brought by all three variations over the original model. Variation 3 yields a slightly higher adjusted R^2 value than that of Variation 2. A closer look at the results shows that the interaction of the indicator variable, VACHOME, with the variable representing the development index, DVLPNDX, is responsible for this improvement. A possible correlation between the actual number of vacation homes and the development index may explain this result. We represent the number of vacation homes only with an indicator variable in Variation 2, whereas, the interaction of the indicator variable with the development index may have served as a slightly better proxy for the number of vacation homes in Variation 3.

We also solved the model for different geographical regions separately. These results were, in

Table 1
Adjusted R^2 values and coefficients of different regression models[a]

	Model			
	Original	Variation 1[b]	Variation 2[b]	Variation 3[b]
Adj. R^2	0.60	0.88	0.86	0.90
SB	−	−	−	−
DVLPNDX	+	+	+	+
PCTOUR	+	+	+	+
PCGDP	+	+	+	+
CLIM	0	0	0	+
CLIMNEGV	0	0	0	0
CLIMVAR	0	0	−	−
CPI	0	0	0	0
DUMPOS1[c]	*	+	*	*
DUMNEG1[c]	*	−	*	*
VACHOME[d]	*	*	+	0
SBINT[e]	*	*	*	−
PCTOURINT[e]	*	*	*	−
DVLPNDXINT[e]	*	*	*	+
Other interactions	*	*	*	0

[a] − implies that the variable has a significant negative effect. + implies that the variable has a significant positive effect. 0 implies that the variable does not have a significant effect. * implies that the variable is not included in the model.
[b] Variation 1: model having two indicator variables. Variation 2: model having an indicator variable to represent cities with summer houses. Variation 3: model having interaction effects.
[c] DUMPOS1 and DUMNEG1 are the indicator variables representing points plus and minus 1 standard deviation away from the corresponding predicted values, respectively.
[d] VACHOME is the indicator variable representing cities having many vacation homes.
[e] SBINT, PCTOURINT, and DVLPNDXINT are the variables representing interactions of SB, PCTOUR, and DVLPNDX, respectively, with the indicator variable, VACHOME.

general, consistent with the results of the overall model. Using the indicator variables again improved the results substantially for some of the regions in which those variables were justified.

In our application, the order of the data was not important. If one has time series data or some other data where the order is important, it may be possible to further generalize the idea we presented regarding the modeling of residuals. In such cases it would be possible to borrow further tools from statistical process control in determining the data points for which to use indicator variables. One could define different zones (based on standard deviations) around the predicted values and identify cases, for example, where two out of three successive data points fall into Zone A or beyond or where four out of five successive points fall into Zone B or beyond. Such data points would then be candidates to be used with indicator variables in the regression. Other tests developed for control charts can also be applied when the order of the data is important (see, for example, [2] pp. 157–164, for tests on out-of-control conditions).

4. Short-term model

This model was developed to identify the factors that affect beer demand in the short term. The time period was selected as a month so that consumers' reactions to the various changes in a given month could be accounted for. This model allowed us to see the effects of factors such as the price of beer, the relative prices of beer substitutes, the effect of Ramadan (a month in the Islamic calendar during which Muslims are required to fast during daytime), etc.

We first tried to use total beer consumption in Turkey in a given month as our dependent variable. The residuals of the model demonstrated nonconstant variance. Then we made a logarithmic transformation of the dependent variable and it worked well in this case. We next state the variables used in the model. The precise calculation of variables is given in the appendix.

Dependent variable:

$LNMS_t$: Natural logarithm of total beer sales in Turkey in month t.

Independent variables:

$TREND_t = t$, where the first month we consider has $t = 1$. This variable was introduced to represent the natural increase in beer sales (due to population growth, change in consumption habits, etc.).

$RAKI\text{-}DRFT_t$: cost ratio of a liter of Raki (a national, widely consumed hard liquor which we use as a representative of hard liquor) to a liter of draft beer. Our clients suggested that, based on the medium of consumption (places where Raki and beer are both consumed), Raki could be a substitute for draft beer.

$SDCAN\text{-}BCAN_t$: cost ratio of average canned soft drink to canned beer in month t. Our clients suggested that canned soft drinks would be substitutes for canned beer.

$AVGCORP_t$: average corrected beer price in month $t + 1$ to be used as the beer price for month t. The monthly beer sales values we use are the amounts bought by the distributors in those months. We realized after making several runs that the distributors are informed about the price increases beforehand. Therefore, here we expect a relation between the beer sales of month t and price in month $t + 1$. For example, they would stock beer in month t if the price would increase in month $t + 1$. We also assumed that a price increase made in month t goes into effect in month $t + 1$ since the distributors have enough time to stock-up right before the price increase in month t.

$TOUR_t$: tourism measure in month t.

RAM_t: number of days in month t coinciding with Ramadan.

$SEASON_t(i) = 1$ if month t corresponds to i and 0 otherwise where $i = 1, 2, \ldots, 8, 10, 11, 12$. This variable was used to measure the seasonality effect for month t. Seasonality effect for September is not used, which means that all seasonality effects are measured relative to September.

We did not use a separate variable to measure the effect of the climate, thinking that the seasonality effect would capture that. All the price values are real prices that are obtained after accounting for the inflation. The prices were increased several times a year because there was high inflation in Turkey during the period of this study.

4.1. Results

We made the regression run for the period starting with January 1987 and ending with December 1993. As we used the prices by shifting one period we could use the beer sales values from January 1987 (which corresponds to month 1 in our model) to November 1993 (month 83). The regression results show that the model explained about 97% of the monthly sales. The graph of the residuals against the predicted values does not show any undesirable (nonrandom) characteristics.

The month of June does not have a significant effect relative to September, whereas July and August have positive effects and the other months have negative effects compared to September. All these seasonal effects are consistent with what one would expect. The TREND and the price ratio of soft drink to beer has a positive effect on beer consumption as expected. Again as expected, Ramadan has a significant decreasing effect on beer consumption and the beer sales of month t has a significant positive correlation with the beer price of month $t + 1$. That is, beer sales in month t increases in response to an increase in beer price in month $t + 1$ and vice versa. In addition to the seasonal effect of June, the price ratio of Raki to draft beer and the variable corresponding to tourism turned out not to have significant effects. The effect of tourism might have been accounted for in the seasonal variables since a very large percentage of tourist activity (both domestic and foreign) occurs during the summer months.

The results of the model indicate that short-term effects on beer consumption have been captured well and the model can be useful in forecasting

short-term beer sales. Moreover, one can analyze elasticity of beer demand with respect to each of the significant variables.

5. Conclusions

We have presented a medium-term and a short-term regression model for explaining and forecasting beer demand in Turkey that was developed collaboratively with our clients. We also outlined a procedure based on examining the residuals, which uses statistical process control principles and indicator variables to detect and incorporate potential missing variables into the regression model. By identifying the aberrant data points, missing lurking variables (which have important effects on these points) can be more easily determined since one needs to concentrate on finding common factors that have substantial effects on these data points only. Both models explained very large percentages of the variations of the respective dependent variables.

A topic for future research is to develop a formal procedure that would help identify common aspects of data points that are most affected by missing variables. These data points could be examined based on some qualitative information (such as the regions in which the cities are located or the year the data point belongs to in our application) with the help of multivariate statistical techniques (such as factor analysis) or pareto analysis. The results of such a formal analysis may further help disclose the missing variables.

During the project we worked very closely with our clients and they actively participated in the project. This contributed significantly to the success of the project and we benefited from interpreting the results together with our clients. We submitted all the models together with a detailed report to the company and they intend to update the data and keep using the models in the future. We believe that this will be possible because they have been actively involved in all phases of the project.

The modeling of residuals we conducted in this study has been shown to be useful both for the beer application and the two randomly generated problems. We intend to test and further develop this procedure by conducting a large set of controlled experiments on randomly generated problems as future research.

Appendix A

The following data was compiled for the computation of variables used in the regression models.

1. The subscript sets are defined as follows:

$p = 1, \ldots, 5$ (denotes different packages of beer, i.e. 50 cl bottle, 30 cl bottle, 50 cl can, 33 cl can and 50 l draft beer, respectively).

$q = 1, \ldots, 96$ (denotes months starting from January 1987 through December 1994)

$j = 1, \ldots, 15$ (defines production activities, i.e. agriculture and livestock production; forestry; fishing; mining and quarrying; manufacturing industry; electricity, gas and water; construction; trade; transportation and communication; financial institutions; owning of dwellings; business and personal services; (less) imputed bank service charges; government services; import taxes).

$k = 1, \ldots, 3$ (denotes years, 1989 through 1991).

$i = 1, \ldots, 67$ (denotes cities in Turkey in alphabetical order. Cities established after 1989 are considered to be within their old territories).

$m = 1, \ldots, 365$ (denotes days).

$n = 1, \ldots, 30$ (denotes countries, or group of countries including Turkey as $n = 30$).

2. The description of the compiled data are given below:

pc_{pq}:	Current price per l of package p at month q.
q_{pq}:	Amount of package p sold by the company in l in month q.
e_q:	Increase (decrease) rate of the consumer price index at month q. Consumer price index value at the beginning of January 1989 is taken as 100.
c_{jt}:	Value added obtained nationwide by the jth production activity in year t, in current prices (source: [8]).

q_{it}: The proportion of value added of the jth production activity generated by city i (source: [5] Estimates of proportions made for 1986 are the most recent available information. Through discussions with SIS experts, the 1986 estimates are assumed to be stable and are used for 1989–1991).

r_t: GDP deflator for year t (source: SIS 1993 and [9]).

N_{it}: Population of city i in year t (source: [6,7]).

KN_{it}: Population of urban part of city i in year t (source: [6,7]).

$\mathrm{KN18}_{it}$: 18 yr or older population of urban part of city i in year t (source: [7]).

S_{itm}: The average temperature of city i at the mth day of year t. (source: General Directorate of Meteorology, data supplied in electronic medium).

g_{int}: Total number of nights spent in city i in year t by tourists from country n (source: [4]).

b_n: Annual average per capita beer consumption in country n.

ag_{nq}: Total number of nights spent by tourists from country n in month q (source: [4]).

TS_{it}: Total beer sales (in l) in city i in year t.

MS_q: Beer sales (in l) in Turkey in month q.

CPI_{it}: Consumer price index in city i in year t (source: [8,9]).

$\mathrm{DVLPNDX}_{it}$: Urban development index for city i in year t (source: [1]).

pr_q: Current price per l of Raki at month q.

ps_q: Current price per l of canned soft drink at month q.

3. The variables used in the regression models are computed using the following formulas:

$$*\mathrm{PCS}_{it} = \frac{\mathrm{TS}_{it}}{1000\,\mathrm{KN18}_{it}}, \quad i = 1, \ldots, 40, \quad t = 1, \ldots, 3,$$

Cities are combined and there are 40 data points, i.e. 40 cities or city groups. $\mathrm{KN18}_{it}$ terms for 1989 are estimated from 1985 and 1990 General Population Census results using a natural increase formula. We collaborated with experts from SIS for the estimation of 18 yr or older urban population. 1991 estimates were taken directly from SIS.

$$*\mathrm{PCGDP}_{it} = \frac{\mathrm{GDP}}{10^9 N_{it}\,\mathrm{RP}_t}, \quad i = 1, \ldots, 40, \quad t = 1, \ldots, 3,$$

where

$$\mathrm{GDP}_{it} = \sum_{j=1}^{15} q_{it} C_{jt} \frac{1}{\sum_{m=1}^{t}(1 + r_t)},$$

$$\mathrm{RP}_t = \frac{\sum_{p=1}^{5}\sum_{q=12(t-1)+25}^{24+12t} \mathrm{pc}_{pq} q_{pq} / \prod_{s=1}^{q}(1 + e_S)}{\sum_{p=1}^{5}\sum_{q=12(t-1)+25}^{24+12t} p_{pq}}.$$

Note that GDP_{it} is the Gross Domestic Product in city (or city group) i in year t and RP_{it} is the real price computed as the weighted average l price of different packages.

$$*\mathrm{PCTOUR}_{it} = \frac{\sum_{n=1}^{30} g_{int} b_n / 365}{\mathrm{KN18}_{it}}, \quad i = 1, \ldots, 40, \quad t = 1, \ldots, 3,$$

$$*\mathrm{CLIM}_{it} = \sum_{m=1}^{365} f(S_{itm}), \quad i = 1, \ldots, 40, \quad t = 1, \ldots, 3,$$

where $f(.)$ is a function defined in three regions (increasing at an exponential rate, constant and decreasing at an exponential rate).

The aim in using $f(.)$ is to give more weight to an intermediate range of temperatures with respect to beer consumption.

$$*\mathrm{CLIMVAR}_{it} = \mathrm{VAR}_{it}^{+} - \mathrm{VAR}_{it}^{-}, \quad i = 1, \ldots, 40, \quad t = 1, \ldots, 3,$$

where

$$\mathrm{VAR}_{it}^{+} = \sum_{q=12(t-1)+25}^{24+12t} \sum_{m=1}^{n_q} [(S_{itm} - \bar{S}_{iq})^{+}]^2 / n_q,$$

$$\mathrm{VAR}_{it}^{-} = \sum_{p=12(t-1)+25}^{24+12t} \sum_{m=1}^{n_q} [(\bar{S}_{iq} - S_{itm})^{+}]^2 / n_q,$$

where $\bar{S}_{iq}$ is the average temperature in city i in month q, n_q the number of days in month q and $(x)^+ = \max(0, x)$. Note that VAR_{it}^+ and VAR_{it}^- denote the sum of squared values of daily temperatures above the monthly averages and below the monthly averages, respectively.

*$\mathrm{CLIMMNEGD}_{it} = \mathrm{VAR}_{it}^-$ (as defined above),

$i = 1, \ldots, 40, \quad t = 1, \ldots, 3,$

*$\mathrm{LNMS}_t = \ln(\mathrm{MS}_t), \quad t = 1, \ldots, 96,$

*$\mathrm{RAKI} - \mathrm{DRFT}_t = \mathrm{pr}_t/\mathrm{pc}_{5t}, \quad t = 1, \ldots, 96,$

*$\mathrm{SDCAN} - \mathrm{BCAN}_t = \mathrm{ps}_t/\mathrm{pcc}_t, \quad t = 1, \ldots, 96,$

where pcc_t is the average current price for a l of canned beer, computed as the weighted average of pc_{2t} and pc_{3t}.

*$\mathrm{TOUR}_t = \sum_{n=1}^{30} \mathrm{ag}_{nt} b_n, \quad t = 1, \ldots, 96.$

References

[1] H. Akder, UNDP's human development report and Turkey country profile. Report on The First National Human Development Conference, Ankara, 1992, pp. 11–25.
[2] R.E. DeVor, T. Chang, J.W. Sutherland, Statistical Quality Design and Control, MacMillan, New York, 1992.
[3] B.J. Mandel, The regression control chart, Journal of Quality Technology 1 (1969) 1–9.
[4] Ministry of Tourism Publication, Bulletin of Accommodation Statistics, General Directorate of Investments, Department of Research and Evaluation, Turkey, 1994.
[5] E. Özötün, Türkiye Gayri Safi Yurt Içi Hasilasinin Iller Itibariyle Dagilimi, 1979–1986, ISO Yayin No: 1988/8, 1988.
[6] SIS (State Institute of Statistics) Publication, 1985 General Population Census, Turkey, 1985.
[7] SIS (State Institute of Statistics) Publication, 1990 General Population Census, Turkey, 1990.
[8] SIS (State Institute of Statistics) Publication, Statistical Yearbook of Turkey – 1993, 1994a.
[9] SIS (State Institute of Statistics) Publication, Turkish Economy: Statistics and Analysis – June/July 1994, Turkey, 1994b.