

An executive summary for managers and executive readers can be found at the end of this article



Steps in forecasting with seasonal regression: a case study from the carbonated soft drink market

Albert Caruana

Professor of Marketing, Department of Marketing, ESC-Toulouse, Toulouse, France

Keywords Forecasting, Statistical forecasting, Soft drinks industry, Seasonal trends

Abstract Forecasting enables the efficient utilisation of a firm's resources. There are various types of forecasting models that can be built. Illustrates the steps involved in building a forecasting model utilising seasonal regression with a practical example. The model obtained for the carbonated soft drink brand under consideration estimates a growth rate of 3,568 units per month during the last five years and identifies the seasonal effect during each month of the year. The model also computes the cannibalisation effect that the introduction of a brand extension has had. The development of such models can provide a useful input to both marketing and operations planning.

Introduction

Forecasting is important to firms because it can help ensure that effective use is made of resources. It can be an important aid in identifying trends in sales and the purchase of raw materials in the correct amounts. There are a number of forecasting techniques or models that are available to management and the choice of technique requires a number of considerations. If management believes that the future facing their firm is predictable or fairly predictable then statistical forecasting is a useful tool. If, on the other hand, an organisation faces a very turbulent environment where the future is mostly unpredictable or wholly uncertain, then there is little point in attempting to utilise statistical techniques to forecast the future. Some qualitative forecasting techniques have been suggested for a mostly unpredictable future scenario (Fahey and Randall, 1998). However, in a wholly uncertain scenario the best a firm can do is to have a structure that is responsive and adaptable enabling it to meet the expected market turbulence.

The two main groups of forecasting techniques that can address an environment that is predictable or fairly predictable are time series and causal techniques (Aiken and West, 1991; Seber, 1977; Weisberg, 1985). Time series represent a group of techniques most associated with predictable futures and include Regression, Decomposition and the various Adaptive methods. With such techniques one essentially seeks to identify patterns in the data over time and moves to project the established patterns into the future. However, in utilising the identified pattern for forecasting it must be stressed that any resultant forecast assumes that "what has happened in the past will continue to happen in the future" – the future is predictable. If for any reason this basic assumption is violated whether as a result of external or internal changes (e.g. the firm intends to launch a massive advertising campaign) the accuracy of the forecast becomes very questionable. With a time series the variable of interest, which is often sales, is considered over time. Clearly, brand sales patterns are not the result of

The research register for this journal is available at
http://www.mcbup.com/research_registers
 The current issue and full text archive of this journal is available at
<http://www.emerald-library.com/ft>

Four main elements

time but could be caused by various marketing activities, such as: advertising, sales promotion, more salespersons and so on. This is a causal model and as the name implies the attempt is to determine the cause and effect relationship that exists between a set of variables. Causal modelling encompasses a variety of techniques that include Linear programming, simulation, and stepwise multiple regression to mention just a few. One of the main problems faced in forecasting using causal models is often the difficulty in identifying suitable leading indicators. For example, if one was trying to forecast the size of the building industry in the months to come, it is likely that the number of permits issued now can probably act as a suitable leading indicator. This is because the issue of the permit precedes the actual buildings (what you are trying to predict) by a few months. It is not always possible to build an acceptable causal model (Sharma, 1999). There may be a variety of reasons for this. These include an inability to identify leading indicators or simply that the data for the leading indicator are not available. In such circumstances one of the time series techniques, particularly decomposition, can be a useful alternative.

Any time series observation consists of four main elements: a seasonality effect, a trend effect, a cycle effect and residual error. Since the cycle effect is often a long-term effect it is often treated as part of the residual "error". Modern software makes it possible to build a model that simultaneously allows the evaluation of the main components of a time series in terms of trend and seasonal effects (cf. Lim and McAleer, 2000; Proietti, 2000). This paper looks at a brand in the carbonated soft drink market to illustrate the use of the seasonal regression technique and its refinement with weighted least squares regression. A five-step procedure is used to estimate the coefficients of the independent variables that also include the introduction of a brand extension. The estimates of the coefficients can be used to forecast sales in the next time periods.

Five-step procedure

Step 1 – Collecting and inspecting the data

Monthly data starting from January 1995 were collected for a carbonated soft drink brand whose time series is shown in Figure 1.

Step 2 – Building the model

It is possible to integrate the effect of seasonal movement in a regression model by incorporating 11 dummy variables for 11 of the 12 months. The twelfth

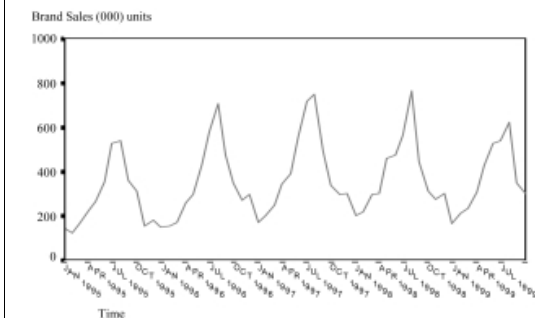


Figure 1. Times series showing brand sales

Outliers

month is reserved as a baseline for comparison. If one uses all 12 months, the twelfth month provides no information that one could not figure out from the first 11. A monthly trend variable will be used. It is clear from Figure 1 that the seasonal effect does not have a more pronounced effect with time. An additive model rather than a multiplicative model can be used. Using the trend variable, the 11 seasonal dummy variables, as well as the effect of the introduction of the improved brand modelled with a dummy variable, it is possible to analyse the time series for the brand.

A first regression analysis (in Table I) indicated the presence of two outliers. The first was item 44 (for August 1998 when a major competitor was out of stock) and item 57 (when a major competitor was running an aggressive sales promotion campaign). Outliers can have a disproportionate influence on trend estimates (Bates *et al.*, 1999; Rousseeuw and Leroy, 1987). Significance tests on regression coefficients depend on the assumption of normally distributed residuals and hence these are also sensitive to outliers. To overcome this problem these two observations were replaced using linear interpolation.

A number of points can be noted from Table I. First, the value of the adjusted R^2 at 0.925 shows that over 90 percent of the variation in the brand sales data is predicted by this model, even when adjusting for the number of variables and months of data. Prior to linear interpolation for the two outliers, adjusted R^2 stood at 0.919. Second, the coefficient of the trend variable shows that the brand has had an upward sales trend of 4,140 units per month over the period being considered. The relative t statistic shows this to be statistically significant at the 99 percent level ($p < 0.01$). Third, the negative coefficient for the improved brand (CANB) that was launched shows that this has cannibalised the sales of the main brand to the tune of over 118,000 units. The

Multiple R	0.971
R square	0.942
Adjusted R square	0.925
Standard error	43.551

Analysis of variance

	DF	Sum of squares	Mean square	F	Sig. F
Regression	13	1,358,103.91	104,469.53	55.08	0.000
Residual	44	83,455.90	1,896.72		

Variables in the equation

Variable	B	SE B	Beta	T	Sig. T
TREND	4.14	5.61	0.44	7.38	0.000
CANB	-118.35	20.93	-0.34	-5.68	0.000
JAN	-89.33	29.29	-0.16	-3.05	0.004
FEB	-76.72	29.26	-0.14	-2.62	0.012
MAR	-40.48	29.24	-0.07	-1.38	0.173
APR	17.31	29.23	0.03	0.59	0.557
MAY	94.82	29.22	0.17	3.25	0.002
JUN	216.94	29.38	0.39	7.38	0.000
JUL	329.484	29.34	0.59	11.23	0.000
AUG	363.654	29.31	0.65	12.41	0.000
SEP	181.07	29.29	0.32	6.18	0.000
OCT	53.70	29.28	0.10	1.83	0.073
NOV	-18.46	30.80	-0.03	-0.60	0.552
Constant	176.87	25.43		6.96	0.000

Table I. Multiple regression after linear interpolation of outliers for the brand

Residual and predicted values

effect is also statistically significant at the 99 percent level. Fourth, each of the dummy month variables shows the seasonal effect of that month compared to December, the omitted month. All seasonal effects are significant at the 95 percent level ($p < 0.05$) compared to December except for March, April and October and November. Finally, the constant term with a value of 176,873 units is the predicted sales of the main line brand at the beginning of the time series (January 1995), after removing the seasonal factors.

Step 3 – Residual analysis

The residual analysis for the regression indicates that item 44 still comes out as an outlier with a residual value of -97.73. This is reflected in the slight departure from normality evidenced on the right hand side of the histogram of the standardised residuals (Figure 2) and in the normal probability plot where deviations from the diagonal can be observed (Figure 3). For tests of significance of regression coefficients to be valid the assumption of normally distributed residuals must hold (Draper and Smith, 1966). This is not the case here.

The scatterplot in Figure 4 compares the residuals on the vertical axis with the predicted values on the horizontal axis. The plot shows a funnel shape: the variance of the points at the right is more than the variance of the points at the left. The shape of the plot of the residuals with the predicted values indicates that the residuals for observations with high predicted sales have

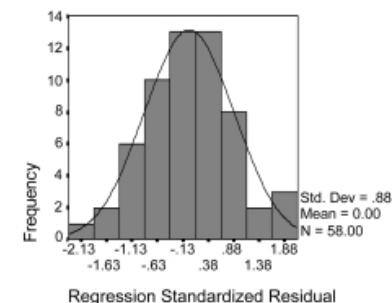


Figure 2. Histogram of standardised residuals

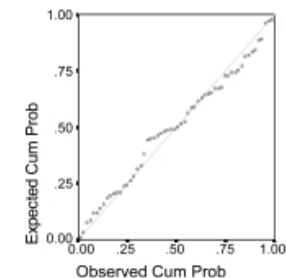


Figure 3. Scatterplot of residuals vs predicted value

Hourglass pattern

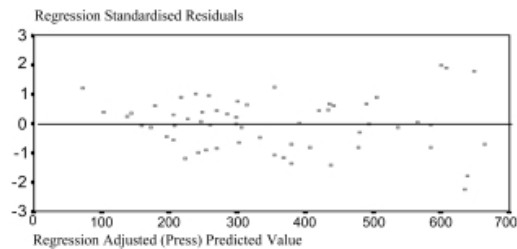


Figure 4. Scatterplot of regression standardised residuals with predicted values

more variance than residuals for observations with low predicted sales. Ordinary regression analysis assumes that residuals have constant variance. This regression model evidently violates this assumption. In other words, the model exhibits heteroscedasticity (Robie and Ryan, 1999).

Step 4 – Investigating heteroscedasticity

The presence of heteroscedasticity is further confirmed by the plot of the residuals against the month of observation in Figure 5. This is not a time series plot; all the Januarys are plotted together, all the Februarys and so on, so that one can evaluate the variance of the residuals in each month. The plot shows an hourglass pattern. The residuals are especially spread out vertically in the summer months. It shows that the error variance differs according to the month of observation. This heteroscedasticity of the residuals violates one of the assumptions of ordinary least squares regression, so that some of the statistical results of the analysis obtained in Table I may not be reliable. To obtain reliable results, weighted least squares regression must be used.

Step 5 – Weighted least squares estimation

The weighted least squares technique of regression performs analysis for observations measured with varying precision. It makes use of the monthly data available but gives more weight to the more precise observations and less weight to the highly variable observations. The procedure estimates the

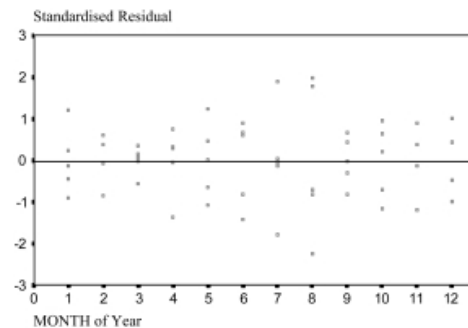


Figure 5. Residual variance by month

Observations

power to which a source variable needs to be raised in order to measure the precision of each observation. The results for the brand using weighted least squares regression are given in Table II.

A number of observations can be made from Table II. First, the value of the adjusted R^2 at 0.941 has improved marginally from the original 0.925 with over 94 percent of the variation in brand sales data predicted by this model. Second, the coefficient of the trend variable shows that the brand has had an upward sales trend of 3,569 not 4,139 units per month over the period being considered. The relative t statistic shows this to be statistically significant at the 99 percent level ($p < 0.01$). Third, the negative coefficient for the improved version of the brand (CANB) shows that this has cannibalised the sales of the main brand to the tune of over 95,000 rather than 118,000 units. The effect is also statistically significant at the 99 percent level. Third, each of the dummy month variables shows the seasonal effect of that month compared to December, the omitted month. Again all seasonal effects are significant at the 95 percent level ($p < 0.05$) compared to December except for March, April and October and November. Finally, the constant term with a value of 188,300 rather than 176,873 units is the predicted sales of the brand at the beginning of the time period (January 1995), after removing the seasonal factors.

Clearly the observations during the summer months had variances that tended to overestimate values resulting from the least squares regression analysis. The weighted least squares estimates are expected to be superior to those obtained with ordinary regression. Figure 6 shows a plot of predictions against residuals and confirms that the heteroscedasticity observed earlier has been dealt with.

Multiple R	0.977
R Square	0.955
Adjusted R Square	0.941
Standard Error	43.90

Analysis of variance:

	DF	Sum of squares	Mean square	F	Sig. F
Regression	13	1,467,819.02	112,909.16	46.78	0.000
Residuals	44	106,206.37	2,413.78		

Variables in the equation

Variable	B	SE B	Beta	T	Sig. T
CANB	-95.54	18.72	-0.32	-5.10	0.000
TREND	3.57	0.49	0.45	7.33	0.000
JAN	-9.10	29.58	-0.18	-3.08	0.004
FEB	-77.86	27.20	-0.20	-2.86	0.006
MAR	-41.05	25.23	-0.13	-1.63	0.111
APR	17.31	29.64	0.04	0.58	0.562
MAY	95.39	30.81	0.18	3.10	0.003
JUN	213.52	32.19	0.36	6.63	0.000
JUL	326.64	35.13	0.47	9.30	0.000
AUG	413.32	40.57	0.48	10.19	0.000
SEP	156.56	27.51	0.39	5.69	0.000
OCT	52.57	30.71	0.10	1.71	0.094
NOV	-19.03	32.24	-0.03	-0.59	0.558
Constant	188.30	25.59		7.36	0.000

Notes:

Log-likelihood Function = -683.64
POWER value = -1.600

Table II. Weighted least square estimation for the brand

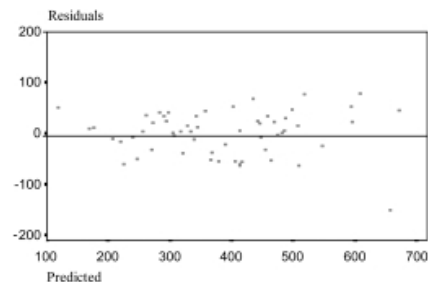


Figure 6. Scatterplot of residuals against predicted values

Conclusion

If no major factors are known to be likely to affect the brand, the above analysis provides coefficients that can be used to generate sales forecasts for the coming few months. The model indicates that the brand is growing at the rate of 3,570 units per month. The introduction of a variety of the original brand has had an annual cannibalisation effect of 95,000 units and the seasonal effect of each month relative to December is also calculated. The development of such models can provide a useful input to both marketing and operations planning.

References

- Aiken, L.S. and West, S.G. (1991), *Multiple Regression: Testing and Interpreting Interactions*, Sage, London.
- Bates, R.A., Holton, E.F. and Burnett, M.F. (1999), "Assessing the impact of influential observations on multiple regression analysis in human resources research", *Human Resource Development Quarterly*, Vol. 10 No. 4, pp. 343-63.
- Draper, N. and Smith, H. (1966), *Applied Regression Analysis*, John Wiley & Sons, New York, NY.
- Fahey, L. and Randall, R.M. (1998), *Learning from the Future: Competitive Foresight Scenarios*, John Wiley & Sons, New York, NY.
- Lim, C. and McAleer, M. (2000), "A seasonal analysis of Asian tourist arrivals to Australia", *Applied Economics*, Vol. 32 No. 4, pp. 99-510.
- Proietti, T. (2000), "Comparing seasonal components for structural time series models", *International Journal of Forecasting*, Vol. 16 No. 2, pp. 247-60.
- Robie, C. and Ryan, A.M. (1999), "Effects of nonlinearity and heteroscedasticity on the validity of conscientiousness in predicting overall job performance", *International Journal of Selection & Assessment*, Vol. 3, pp. 157-69.
- Rousseeuw, P.J. and Leroy, A.M. (1987), *Robust Regression and Outlier Detection*, Wiley & Sons, New York, NY.
- Seber, G.A.F. (1977), *Linear Regression Analysis*, Wiley & Sons, New York, NY.
- Sharma, S. (1999), "The challenge of predicting economic crisis", *Finance & Development*, Vol. 36 No. 2, pp. 40-2.
- Weisberg, S. (1985), *Applied Linear Regression*, Wiley & Sons, New York, NY.

This summary has been provided to allow managers and executives a rapid appreciation of the content of this article. Those with a particular interest in the topic covered may then read the article in toto to take advantage of the more comprehensive description of the research undertaken and its results to get the full benefit of the material present

Executive summary and implications for managers and executives

Better forecasting from the inclusion of seasonal variations

Forecasting is, much of the time, a depressing exercise. It is, as some way once commented, like driving a car blindfold with the navigator looking out the rear window. You can't win. Whatever you do and however well structured your work, you'll still get caught out by, in Harold Macmillan's words, "events, dear thing, events".

But this depressing reality shouldn't stop us from doing the work. It's better to peer ahead into future gloom than to shut our eyes and assume everything will be OK. And, some things are more easily predicted than others. We can use the tools of forecasting to develop the right strategies although we should always remember that the strategic choice still involves a risk our forecast might be wrong (ask the weather man) or our response to the forecast might be wrong. Or both!

Time series versus "cause and effect"

The commonest type of forecast is the straightforward extrapolation of past trends. The problem is that, as the warning on mutual trust ads goes, the past is not necessarily a guide to the future. The fact that it followed a particular trend this year (or even over the last ten years) does not mean that we can be sure of the trend continuing into the future. And, the further ahead we look, the less confidence we can place on the forecast.

Another means of making forecasts is to search for "leading indicators" that suggest change. Certain actions imply subsequent actions – the granting of building permissions suggests the need for builders. But this is not necessarily the case and, given the way in which things interconnect, we may find it difficult to identify leading indicators that provide the right guidance. However, we also have the problem of finding the data we need to make "cause and effect" predictions and have to fall back on qualitative assessment (often called guesswork).

A further complication with forecasting is the extent to which our actions influence the accuracy of the forecast. Given that we will do something, that action must influence the market and affect the way in which we perform to prediction. In the case of a straightforward sale prediction for a subscription business, the forecast of sales depends on the promotional activity. If we don't send out renewal notices and seek to recruit new customers then we will not meet our forecast since that forecast assumes this activity.

Even with bigger and grander forecasts – on national economic performance, for example – the actions of the main player (i.e. the Government) will influence the way in which the economy performs. We can assume that the choices made about levels of tax, rates of interest and borrowing will influence the economy, even if we are not entirely sure about how or how much.

Doing the forecasting

Having depressed you all with my comments about the problems with forecasting, we should consider how to do that forecasting. Caruana points out that time series forecasting encompasses four main elements – a trend effect, a seasonality effect, a cycle effect and a residual error. He goes on to describe how to incorporate the effect of seasonal movements within a forecasting model using regression techniques.

For many businesses this is a welcome addition to forecasting tools since seasonality is a primary factor driving sales. It's not just ice cream makers who sell more at particular times of the year and many businesses know their sales are seasonal but find it difficult to undertake a forecast that encompasses both the underlying trend and the seasonal variation.

Caruana's model focuses on relatively short-term forecasting where the trend effect is less pronounced. If the trend is for a 2 percent annual increase in the market size, then the monthly increase will be very small and could be lost in the variations resulting from seasonality or the residual error.

However, the approach is robust, which suggests that longer-term extrapolations will be valid. We should be able to predict sales on a month-by-month basis taking into account seasonal variations. This assumes that the seasonal variation does not become more pronounced as the market grows and that other variables such as the firm's promotional activity do not skew results. If we shift from an even spread of promotion across the year to targeting peaks, the result may be increased sales at seasonal peaks while seasonal troughs remain static.

Nevertheless, good forecasting should be able to account for the effects of different marketing strategies on overall sales. What emerges from the forecasting process will provide information for the firm to make the right strategic decisions about advertising and promotion. In addition, we can see from Caruana's example how the effects of new products or product extensions can be incorporated into a forecast.

(A précis of the article "Steps in forecasting with seasonal regression: a case study from the carbonated soft drink market". Supplied by Marketing Consultants for MCB University Press.)