Finding Drivers of Biodiversity Offering Sales

1 Introduction

An autoregression model can quantify the relationship between certain drivers and sales potential. In particular, such a model fitted to data can help identify what customer groups are driving sales growth where these groups are defined by different combinations of *gender, age, income, home ownership*, and *education level*.

Another important metric for campaign decision making is the probability of a repeat customer.

Both of these models can be constructed and fitted to data as described in the following sections.

2 Modeling Sales Potential of a New Product

2.1 Measuring sales potential

Letting sales at time t be $S(t)$, one measurable way to quantify sales potential is with the growth rate through time of the biodiversity offering's sales: $SGR_{it} = dS(t)/dt$, i.e., the instantaneous change in sales at time t.

2.2 Predictor variables

1. $\texttt{gender}_i$: gender of customer i: male, female, other.
2. $\texttt{age}_i$: age class of customer i: teen, twenties, thirties, forties, fifties, sixties, senior.
3. $\texttt{income}_i$: household income in USD of customer i.
4. $\texttt{ownhome}_i$: home ownership status of customer i. Takes on the values *renting*, or *own home*.
5. $\texttt{education}_i$: education class of customer i: up through secondary, up through post-secondary, graduate_degree.
6. $\texttt{channel}_i$: the advertising channel that customer i used to learn about the offering. Takes on the values TV, radio, Facebook, X, Instagram, WhatsApp, and print magazines.

2.3 Model

Letting the time points be $t_1, \ldots, t_J$, the model is

$$\begin{aligned} SGR_{t_j,k,l,m,n,o} &= \beta_0 + t_j\beta_{t_j} + \beta_{\text{gender}_k} + \beta_{\text{age}_l} + \beta_{\text{income}_m} + \beta_{\text{ownhome}_n} \\ &\quad + \beta_{\text{education}_o} + Z_{t_j} \end{aligned} \tag{1}$$

where

$$Z_{t_j} = -\sum_{k=1}^{p} \phi_k Z_{t_{j-k}} + \epsilon_{t_j}, \tag{2}$$

and

$$SGR_{t_j,k,l,m,n,o} = \frac{ST_{t_j,k,l,m,n,o} - ST_{t_{j-1},k,l,m,n,o}}{t_j - t_{j-1}}. \tag{3}$$

The computed variable, $ST_{t_j,k,l,m,n,o}$ is total sales during time interval $\{t_{j-1},\ t_j\}$ across all customers who are in gender class k, age class l, income class m, ownhome class n, and education class o.

Channel choice is not a predictor variable in this model because if it were, independence across groups could not be guaranteed. This is because a customer may use different channels at different times, and hence, two goups may contain the same customer at two or more different time points.

2.4 Data requirements

A data set is needed to fit this model. Specifically, observations are needed on the set of variables, $\{S_{i,t_i},\ \text{gender}_i,\ \text{age}_i,\ \text{income}_i,\ \text{ownhome}_i,\ \text{education}_i\}$ on customers $i = 1, \ldots, n_c$, at times $t_i = 1, \ldots, T_i$.

3 A Model of Repeat Purchasing

Let r_{i,t_i} be h if customer i used channel h to help them reach their final decision to go ahead and purchase the biodiversity offering at time t_i where $h = 1, \ldots, H$ with $H > 1$. If the biodiversity offering is an insurance policy, it may be able to be purchased for a one-time annual fee. In this case, if a customer purchased an annual insurance premium and has not cancelled the policy up through time t_i, assume that the customer used the same channel at time t_i as they did when they purchased the policy.

Channel choice for making biodiversity offering purchase decisions through time is modeled with a multinomial, generalized logit, time series model:

$$\begin{aligned} \text{logit}(h_{i,t_i}) &= \beta_{0,h} + t_i\beta_{t_i,h} + \beta_{\text{gender}_{i,h}} + \beta_{\text{age}_{i,h}} + \beta_{\text{income}_{i,h}} + \beta_{\text{ownhome}_{i,h}} \\ &\quad + \beta_{\text{education}_{i,h}}. \end{aligned} \tag{4}$$

The SAS code file, `channels.sas` at `../software/index.html` fits this *autoregressive logistic regression* model to a small, hypothetical data set that is included in that file.

3.1 Data requirements

A data set is needed to fit this model. Specifically, observations are needed on customers who have completed a biodiversity offering purchase. For each of these customers, observations are needed on the following set of variables:

$$\{S_{i,t_i},\ \text{channel}_{i,t_i},\ \text{gender}_i,\ \text{age}_i,\ \text{income}_i,\ \text{ownhome}_i,\ \text{education}_i\}\,.$$

4 Notes

These models are predictive and hence, once fitted to data, can be used to predict conditions that will lead to growth in biodiversity offering revenue in the future.

If the number of observations is less than the number of model parameters in the model, modern methods will be needed to fit these models to data and to test hypotheses with them. The authors of `guerrier_bias_correction.pdf` give one such method. This situation may arise when fitting the generalized logit model of channel choice because such a model's parameter count grows multiplicatively with H.

Appendix: Autoregression

Also called regression with autocorrelated or autoregressive errors.

1. SAS `proc autoreg` fits the model:

$$Y_t = \beta_0 + \sum_{i=1}^{K} \beta_i x_{it} + Z_t \tag{5}$$

   where

$$Z_t = -\sum_{i=1}^{p} \phi_i Z_{t-i} + \epsilon_t \tag{6}$$

   and ϵ_t is i.i.d. $N(0,\, \sigma^2)$, i.e., a *white noise* process.

2. Alway use the Maximum Likelihood (ml) estimation method.

3. `proc autoreg` can test for the three assumptions that are made when this model is employed:

   (a) $\mathrm{Var}[\epsilon_t] = \sigma^2$ for all t (homoscedasticity of white noise variance)

   (b) Autocorrelation of the white noise process is zero at any lag.

   (c) $\epsilon_t \equiv$ i.i.d. $N(0,\, \sigma^2)$.

4. These assumptions are listed in their order of importance.

5. Heteroscedasticity is tested for with two tests: the Portmanteau Q test and the Engle-Lagrange test.

6. The Durbin-Watson test is applied to the residuals *after* correcting for autocorrelation if `nlag` is nonzero. Doing so is a bit theoretically-compromised because $\epsilon_t^* = y_t - y_t^*$ where y_t^* is the *predicted* value of Y_t.

7. This not-so-legitimate fact is a bit hard to find in the SAS documentation but is mentioned in their Money Demand example.

8. Apparently, the normality of ϵ_t is tested for with the Bera-Jarque test.

4.1 Prediction

1. Let Y_t^* be the predicted value at t, and let Z_t^* be the predicted autoregressive error value at t.

2. In `proc autoreg`'s output statement, Y_t^* values are produced with `p=` set to a user-defined name, and Z_t^* values are produced with `rm=` set to a user-defined name.

3. `proc autoreg` refers to Z_t^* as the "residuals from the structural prediction." The structural component of the model is the trend component: $\beta_0 + \sum_{i=1}^{K} \beta_i x_{it}$.

4. A prediction at time t is computed with

$$Y_t^* = \hat{\beta}_0 + \sum_{i=1}^{K} \hat{\beta}_i x_{it} + Z_t^* \tag{7}$$

where

$$Z_t^* = -\sum_{i=1}^{p} \hat{\phi}_i Z_{t-i}^*. \tag{8}$$