ISE 533: Integrative Analytics, Spring 2020

LEO-Wyndor Draft

Group 1:

Xinjie Sun

Xinrui Yan

Yiwen Cao

Zelong Chen

Abstract 3

Goal and Scope of Project 3

Overview of Models/Algorithms 4

- Linear Regression 4
- Deterministic LP Formulation 4
- Stochastic Decomposition (SD) LP Formulation 5
- Sample Average Approximation Formulation 6
- Validation 6

Data Sources 7

- Parameters Assumption 7
- Advertising Dataset 7

Discussion of Results 8

Future Work 9

Abstract

In this project we seek to integrate Marketing predictions with Production optimizations into a single model. Normally Production optimization optimizes based on the predicted sales from Marketing predictions. However, it does not consider the uncertainty that arises from these predictions. By integrating both Marketing predictions with Production optimizations into a single model, we can then incorporate the uncertainty from Marketing predictions into Production optimizations. Thus, decrease the bias in optimization and optimize to the true optimal value. In addition, we will include a validation process that will confirm the decisions made by our models.

Goal and Scope of Project

We look to compare the performance and results from three models: Deterministic, Sample Average Approximation (SAA), and Stochastic Decomposition (SD). We first randomly split the data into 50-50 training and validation sets. Then the amount of total sales could be predicted by linear regression based on the expenditures of TV ads and Radio ads from the training set. Hence, the error terms of training set as well as validation set could be calculated. We applied Chi-square test to the error terms of the training set and validation set, the result showed that they follow the same distribution. The Q-Q plot provided the same result and we dropped several outliers based on that plot. After that, we solved the deterministic model with a production constraint which doesn't contain the error term, but for SAA and SD models, both of them contain the error term. By taking advantage of the optimal value of expenditures on TV ads and Radio ads, we were able to calculate the confidence interval of the validation set and verify whether the optimal value of profits are within the confidence interval. Last but not least, we compare the result of three different methods to figure out which one has the best quality.

Assumptions

- The value of expenditures on TV ads and Radio ads should be within the range of upper bound and lower bound which obtained from the data set.
- The error terms of the training set and the validation set are independent random variables and approximately normally distributed.
- The coefficients of the linear regression are quite stable and the uncertainty is caused by the error term when the degrees of freedom exceed 60.
- All the outliers found in the Q-Q plot were dropped.

Overview of Models/Algorithms

Linear Regression

$$w_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon_i$$

As is shown above, the formulation of the linear regression could be divided into two parts. The first part represents the estimated mean of total sales according to the expenditures on TV ads and radio ads. Another component is the error term which represents the uncertainty of the sales. We made the assumption that the error terms are independent random variables and approximately normally distributed.

Deterministic LP Formulation

Below shows the formulation for the Marketing (1a-e) and Production (2a-f) linear program.

$$\begin{array}{llll}
Max & -0.5x_1 - 0.2x_2 + \mathbb{E}[\text{Profit}(\tilde{\omega})] & & (1a) \\
s.t. & x_1 + x_2 & \leq 200 & (1b) \\
 & x_1 - 0.2x_2 & \geq 0 & (1c) \\
 & -x_1 + 0.7x_2 & \geq 0 & (1d) \\
 & L_1 \leq x_1 \leq U_1, \quad L_2 \leq x_2 \leq U_2 & & (1e)
\end{array}$$

$$\begin{array}{llll}
Profit(\omega) = Max & 3y_A + 5y_B & & (2a) \\
s.t. & y_A & \leq 8 & (2b) \\
 & 2y_B & \leq 24 & (2c) \\
 & 3y_A + 2y_B & \leq 36 & (2d) \\
 & y_A + y_B & \leq \omega & (2e) \\
 & y_A, y_B & \geq 0 & (2f)
\end{array}$$

In the Deterministic LP formulation (3a-f) we will be combining both Marketing and Production LP into one with the addition of the Linear Regression. Linear Regression replaces the ω in the Production LP, as the Linear Regression uses advertising allocation of TV ads (x_1) and Radio ads (x_2) to predict the potential sales ω.

$$\begin{array}{llr}
Max\ -0.5x_1 - 0.2x_2 + 3y_A + 5y_B & & (3a) \\
s.t. \quad x_1 + x_2 & \leq 200 & (3b) \\
x_1 - 0.2x_2 & \geq 0 & (3c) \\
-x_1 + 0.7x_2 & \geq 0 & (3d) \\
y_A & \leq 8 & (3e) \\
2y_B & \leq 24 & (3f) \\
3y_A + 2y_B & \leq 36 & (3g) \\
-\beta_1 x_1 - \beta_2 x_2 + y_A + y_B & \leq \beta_0 & (3h) \\
y_A, y_B & \geq 0 & (3i) \\
L_1 \leq x_1 \leq U_1, \quad L_2 \leq x_2 \leq U_2 & & (3j)
\end{array}$$

Note: In the Deterministic LP Formulation, uncertainty from the Linear Regression model is not included. Thus, the LP will optimize towards the predicted values of Linear Regression while disregarding the uncertainty.

Stochastic Decomposition (SD) LP Formulation

Below shows the formulation of Stochastic Decomposition which includes the error term.

$$\begin{array}{llr}
Max\ -0.5x_1 - 0.2x_2 + 3y_A + 5y_B & & (4a) \\
s.t. \quad x_1 + x_2 & \leq 200 & (4b) \\
x_1 - 0.2x_2 & \geq 0 & (4c) \\
-x_1 + 0.7x_2 & \geq 0 & (4d) \\
y_A & \leq 8 & (4e) \\
2y_B & \leq 24 & (4f) \\
3y_A + 2y_B & \leq 36 & (4g) \\
-\beta_1 x_1 - \beta_2 x_2 + y_A + y_B & \leq \beta_0 + \varepsilon_{ti} & (4h) \\
y_A, y_B & \geq 0 & (4i) \\
L_1 \leq x_1 \leq U_1, \quad L_2 \leq x_2 \leq U_2 & & (4j)
\end{array}$$

In the Stochastic Decomposition Formulation, we included the error terms which represents the uncertainty. What's more, thirty replications were automatically undertaken to obtain low variance estimates of the sample mean. Due to that, the objective value came from the SD solver has the lowest standard deviation and the longest processing time simultaneously.

Sample Average Approximation Formulation

Below shows the formulation of sample average approximation.

$$Max - 0.5x_1 - 0.2x_2 + \frac{1}{N}\sum_{i=1}^{N} 3y_{Ai} + 5y_{Bi} \tag{5a}$$

$$s.t \qquad x_1 + x_2 \leq 200 \tag{5b}$$

$$x_1 - 0.2x_2 \geq 0 \tag{5c}$$

$$-x_1 + 0.7x_2 \geq 0 \tag{5d}$$

$$y_{Ai} \leq 8 \quad i = 1, \ldots N \tag{5e}$$

$$2y_{Bi} \leq 24 \quad i = 1, \ldots N \tag{5f}$$

$$3y_{Ai} + 2y_{Bi} \leq 36 \quad i = 1, \ldots N \tag{5g}$$

$$-\beta_1 x_1 - \beta_2 x_1 + y_{Ai} + y_{Bi} \leq \beta_0 + \varepsilon_{ti} \quad i = 1, \ldots N \tag{5h}$$

$$L_1 \leq x_1 \leq U_1, L_2 \leq x_2 \leq U_2 \tag{5i}$$

$$y_{Ai}, y_{Bi} \geq 0 \quad i = 1, \ldots N \tag{5j}$$

Similar to the stochastic decomposition model, we also included the error term in the formulation of sample average approximation. In this model, we used the average value of finite data points to represent the expected profit which allows us to get a more accurate result compared with the deterministic model. However, we didn't run replications which means it has a higher variance estimate of the sample mean, but takes less time to get the optimal value compared to the stochastic decomposition model.

Validation

Validation is used to validate the decision that was produced by our models using the training data. Using the validation datasets that we can generate the residual error terms, ε_{vi}, by subtracting our predictions using the Linear Regression from the actual sales observation.

$$\varepsilon_{vi} = w_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i}$$

We used Chi-squared test and Q-Q plot to verify that the validation error terms are within the same distribution as the training error terms. We removed any outliers that did not follow the same distribution. Then using the validation error terms and the decision we generated from the training data, we can then create different sales outcomes ω_i .

$$w_i := \beta_0 + \beta_1 \hat{x}_1 + \beta_2 \hat{x}_2 + \varepsilon_{vi}$$

With the sales outcomes ω_i , we will feed it into our Production linear program to generate profits outcomes.

$$\begin{array}{rrll}
Profit(\omega) = Max & 3y_A + 5y_B & & (2a) \\
s.t. & y_A & \leq 8 & (2b) \\
 & 2y_B & \leq 24 & (2c) \\
 & 3y_A + 2y_B & \leq 36 & (2d) \\
 & y_A + y_B & \leq \omega & (2e) \\
 & y_A, y_B & \geq 0 & (2f)
\end{array}$$

Using the profit outcomes from the last step, we can then generate different objective values that simulate different outcomes that can occur under the same decision. With these objective values, we can generate the average and standard deviation to create a 95% confidence interval. Our initial decision will be validated if the initial objective value from the training data falls within this confidence interval.

$$c_1\hat{x}_1 + c_2\hat{x}_2 + \frac{1}{N_v}\text{Profit}\sum_{i\in V}(\omega_i)$$

Data Sources

Parameters Assumption

- Parameters of the optimization models are based on the LEO-Wyndor example in the publication of *Coalescing Data and Decision Sciences for Analytics by Yunxiao Deng, Junyi Liu, Suvrajeet Sen* with some slight modifications

Advertising Dataset

- Advertising dataset that denotes TV, Radio, and Newspaper ad expenditures with associated sales number are found on the cORe website.
- https://core.isrd.isi.edu/chaise/record/#1/Core:Step/RID=W1QY

Discussion of Results

Methodology	x_1	x_2	MPO ($)	MVSAE (1-Fold CV)	MVSAE (5-Fold CV)	Time(s)
Deterministic LP	9.92	49.60	$49,636	$49,403 (±1,050)	$49,403 (±1,110)	0.291
SLP with SAA	9.92	49.60	$49,616	$49,403 (±1,050)	$49,403 (±1,110)	1.157
SLP with SD	9.92	49.60	$49,370	$49,403 (±1,050)	$49,403 (±1,110)	68.453

From the table above, we could figure out that values of first stage variables x_1 and x_2 are the same for all three models. Due to that, in the validation part, the 95% confidence intervals of all three models are identical as well.

The standard deviation of 5-Fold cross validation is a bit higher than that of the 1-Fold cross validation because the sample size of 5-Fold CV is smaller than 1-Fold CV. For all of the three models, all of the three predictions fall within the 95% Confidence Interval, however, there exists slight differences between each other. MPO of deterministic LP and SAA are higher than the estimated mean profit while the outcome of the SD solver is quite close to the estimated mean which means the SD solver provides a more reliable solution.

When it comes to the processing time, obviously the SD solver takes the longest time due to 30 replications. The SAA method also takes more time than the deterministic LP because the number of linear constraints in SAA is about 30 times that of the deterministic LP.

Future Work

Replications

- As suggested in the LEO-Wyndor supplement notes, the method of splitting the dataset into 50-50 training and validation sets could result in a relatively larger variance. As the SD solver provides us with thirty default replications and gives us the optimal values. Although it may take more time to process, estimated mean profit with low variance could be obtained which makes more sense. Hence, it is necessary to carry out replications for deterministic LP and SAA to get solutions of better quality.

Possible modeling adoption for larger dataset

- For the dataset used in this model, SD solver could suffice, though with larger time cost. However, if the need for larger dataset with each production arises, the adoption of LP software like combination of PySP and SD algorithm might be useful since the size of SLP will be large.