**ISE 533 Project 3 Report**

**LEO-Wyndor-2020**

**Advertising and Production Problem**

**Qinhao Chang**

**Tianxiao Yu**

**Shaoqian Chen**

**Wei Li**

**Abstract**

This project focuses on the problem related to the advertising and production of the company WYNDOR Glass. WYNDOR Glass sells two types of doors, type A and type B. They sell these two types of doors based on customer demands. Demands of doors will be guided by advertisement, so advertisement plays an important role in making profit. There are two types of advertisement, TV and Radio. So it is important to decide how much to invest on each type of advertisement.

This project has two goals: decide expenditure of TV and Radio and maximize the profit.

**Introduction**

As described above, in order to decide how much amount to invest on TV and Radio, as well as to maximize the profit, the group builds a 2 stage model. We use x1 and x2 to denote the TV and Radio expenditures. yA and yB to denote the amount of production of type A door and type B door. To build the model, we have several constraints. First one is the total budget, we have 200 thousands of time slots for TV and Radio expenditure. Second, Radio advertisement slots are at least some positive fraction of expenditures in TV slots, the fraction denoted as α, α ∈ (0, 1). Third,due to the range of data in the Advertising data set, we set up the lower and upper bound for both expenditure of TV (x1) and Radio(x2) to make sure the model only includes predictions within the range of data in the advertising data set.

We assume the advertising costs for TV is $500, which denoted as c1, the advertising costs for Radio is $200, which denoted as c2. Total budget is 200000 time slots. Time slots for TV must be greater than 20 percent Time slots for Radio and smaller than 70 percent time slots for Radio.

Furthermore, the production of doors needs to meet the demand by customers.

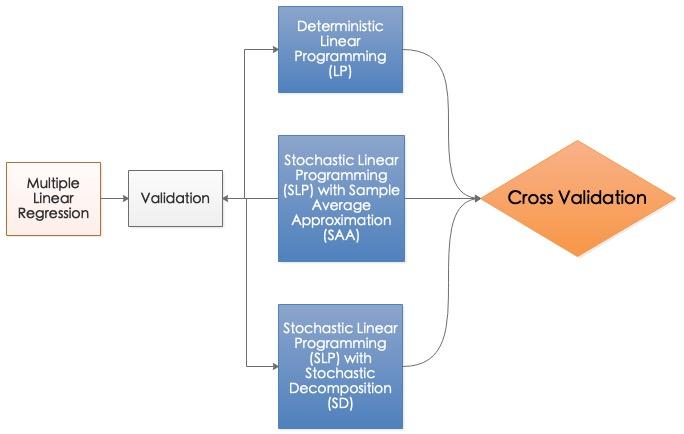
**Overview of Models and Methods**

This section will introduce the models and overview of 3 methods we use to solve this problem:

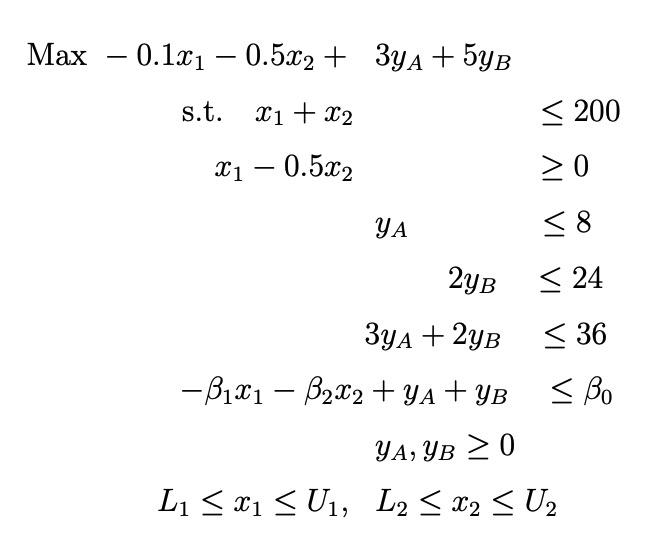
Multiple Linear Regression + Deterministic Linear Program; Multiple Linear Regression + Sample Average Approximation; Multiple Linear Regression + Stochastic Decomposition.

We first build the linear model of the problem. Then we split the advertising data of TV, radio and sales to 50 percent of the training set, 50 percent of the validation set. And use the training set to fit multiple linear regression to get coefficients. In the model, the coefficients are denoted as .

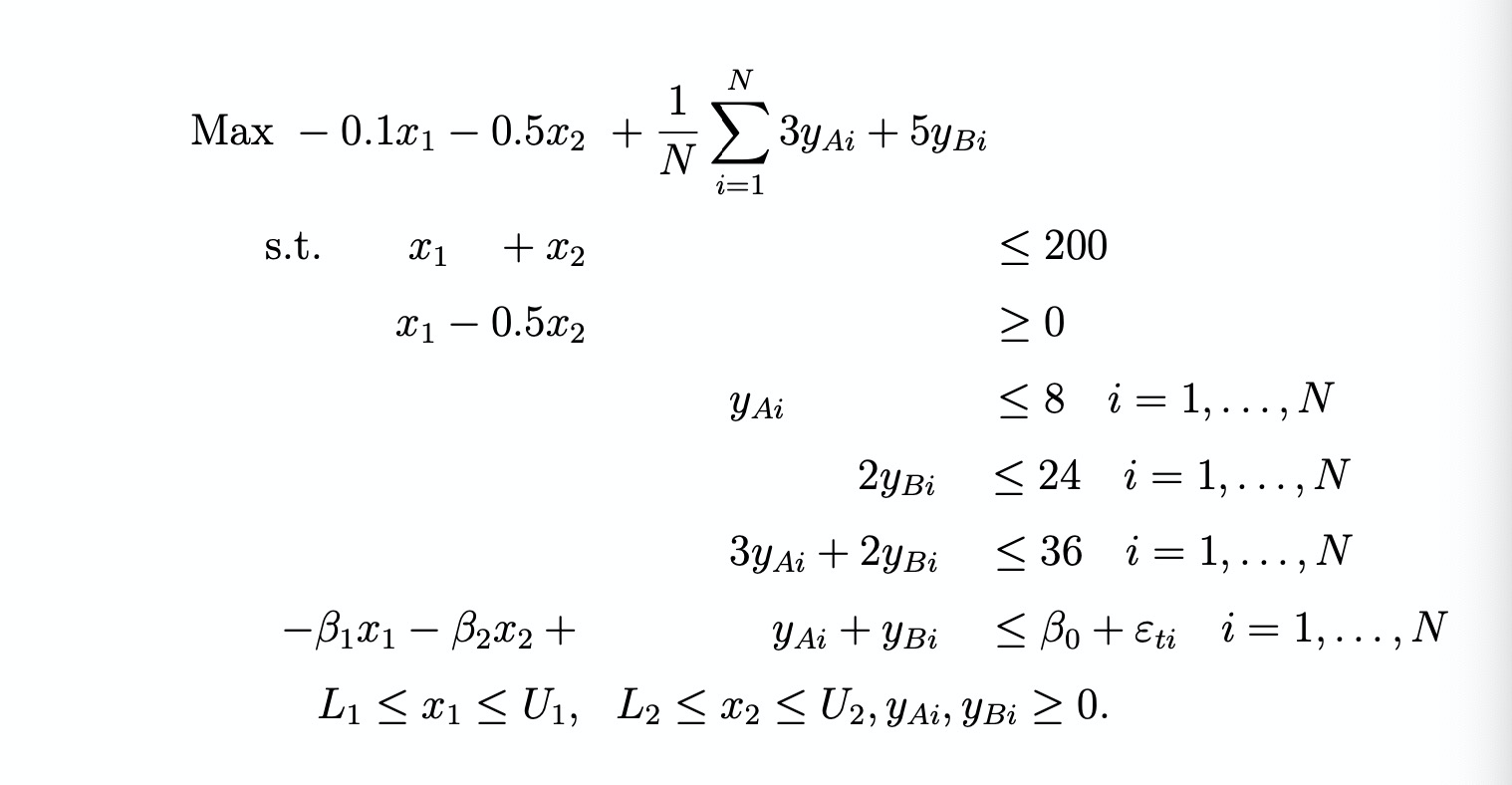
Before using coefficients, we do the validation like 2 sample chi square validation, two sample K-s tests. After validation, wewe write a python code including a cbc solver to solve this linear program to get the value of x1 and x2, as well as the objective value. Then we do the 5-fold cross validation to get the mean objective value and 95 percent Confidence Interval. The general flowchart is shown below.



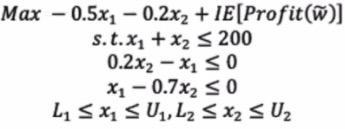
In this project, we treated all numbers as thousands. For example, the cost for TV advertisement is $500, in our model, it will be denoted as $0.5. Based on constraints and the objective,the two stage integrated Linear Programming model is shown below.

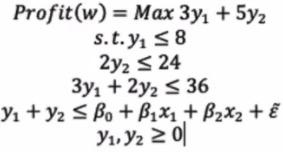


For Sample Average Approximation(SAA) and Stochastic Decomposition(SD), we take the effect of uncertainty in the model. The SAA model is shown below.



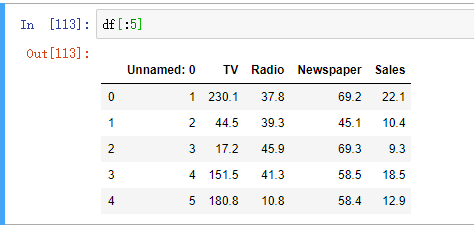
The Model for SD is shown below.

First stage： 

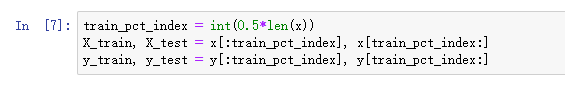
Second stage: 

**Data Sources and Data Science**

The data we use in this project is called ‘Advertising1.csv’. It includes data of expenditures of TV, Radio, Newspaper and Sales. A part of data is shown. In this data, we do not take Newspaper into consideration. TV and Radio are two predictors and Sales are output.

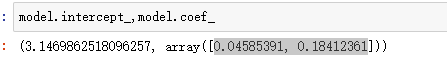


We first split the data into 50% as the training set and 50% as the validation set. That means in each set we have 100 rows of data. We take the first 100 data as training set and the last 100 data as validation set.

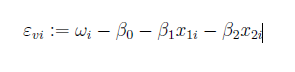


X\_train denotes x1 and x2 in the training set, X\_test denotes x1 and x2 in the validation set, y\_train denotes sales in the training set, y\_test denotes sales in the validation set.

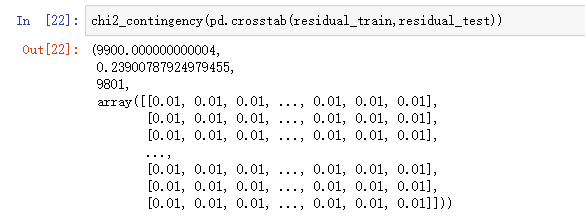
After setting up the training set and the validation set, we fit a linear regression model with TV, Radio, and Sales. The coefficients are 3.1469862518096257,0.04585391, 0.18412361.



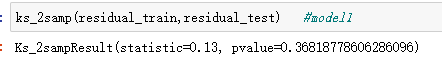
Before using these coefficients, a validation must be done for errors of the training set and the validation set. The validation is to test if errors are from the same distribution. Based on coefficients we get, we calculate the residual error for the training set and validation set by using two formulas.

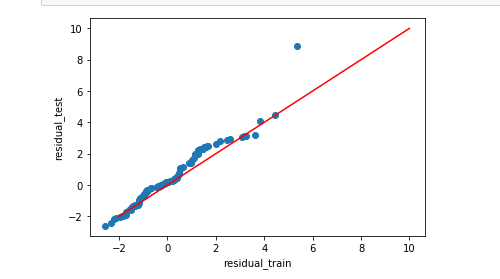
Error is calculated by sales minus the prediction of the model. After we get the error of the validation set and the training set, we perform two sample chi square test and two sample K-S test. The result of two sample chi square test is shown.



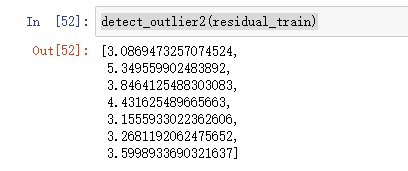
The result of two sample K-S test is shown.



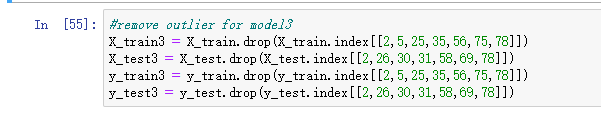
From the result above, the p-value generated by 2 sample chi square test is 0.239 and p-value generated by 2 sample k-s test is 0.36. The p value is larger than 0.05, so we fail to reject the hypothesis that errors are from the same distribution. We then draw a Q-Q plot of error in the training set and error in validation set.



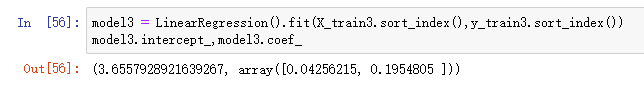
From the Q-Q plot, our group observes that there are outliers. We decide to remove the outlier. We find the outlier in the training set as shown below.



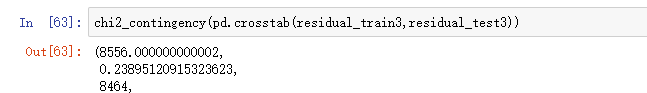
We get the index of outliers which are 2,5,25,35,56,75,78 for the training set and index of outliers which are 2,26,30,31,58,69,78 for the validation set. This means we drop those indices of numbers in the training set and the validation set. The procedure is below.

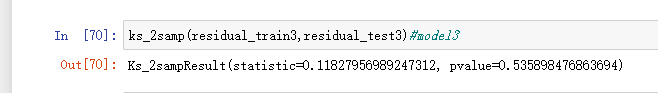


After we drop the outlier using index, we can fit a linear regression model to get coefficients.



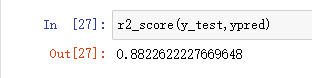
Same as before, we have to make a validation to use these coefficients. We calculate the error in the training set and the validation set and do the same two sample chi square and k-s test.



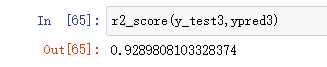


The p value is larger than 0.05 so we fail to reject the null hypothesis that two errors are from the same distribution. We want to know which model is better, so we calculate the R squared value of both models.

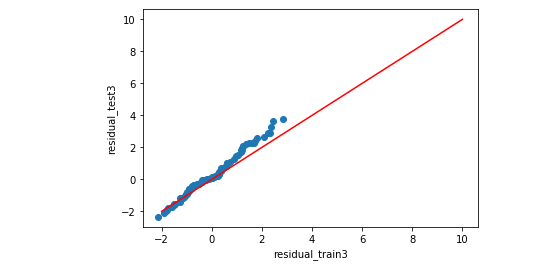
Model with no outliers removed:



Models with 7 outliers removed:



From the result above, we see when outliers are removed, r squared value increases so the model with outliers removed is better than the previous one. We also draw the Q-Q plot after removing outliers.

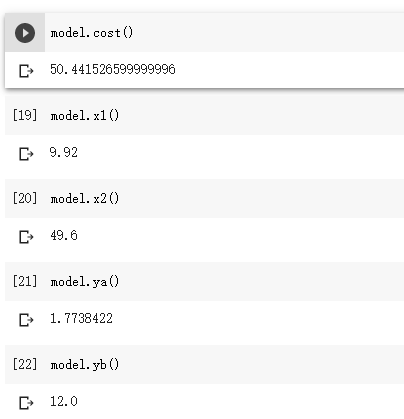


From the graph that scatters fit the line better than before. As a result, our group chooses coefficients 3.6557928921639267, 0.04256215, 0.1954805, corresponding to beta0, beta1 and beta2. There are 93 data point in the training set and 93 data points in the validation set.

**Results and comparisons**

**Deterministic LP**

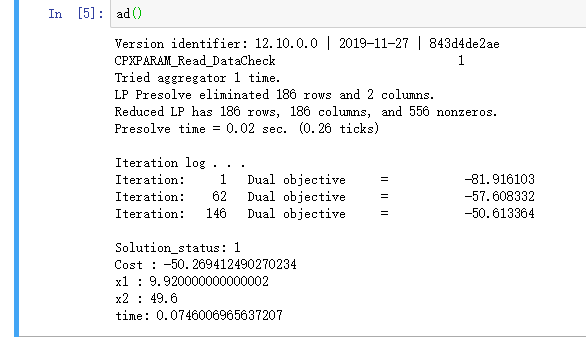
After getting the coefficients, we can solve the linear program we build. The way to solve the linear program is to write python code including ‘cbc’ solver. The result is shown below.



From the result, the objective value of the profit is 50441.52, tv time slots are 9.92, Radio time slots are 49.6. Running time is 0.04second.

**Sample Average Approximation**

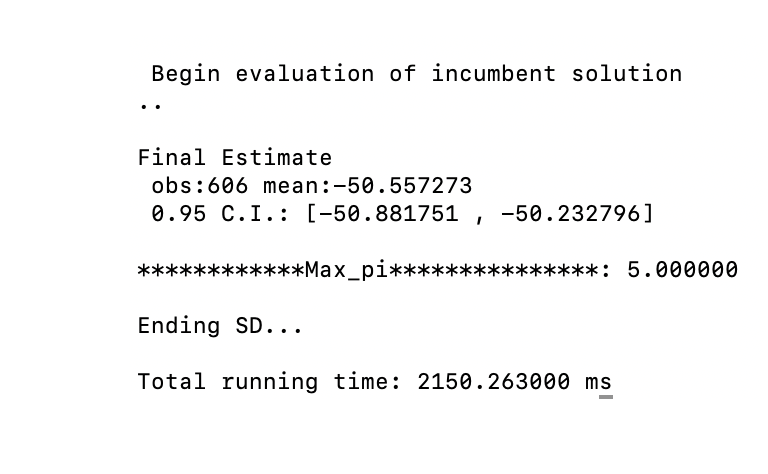
We write python code for SAA to solve the linear program. The result is shown.

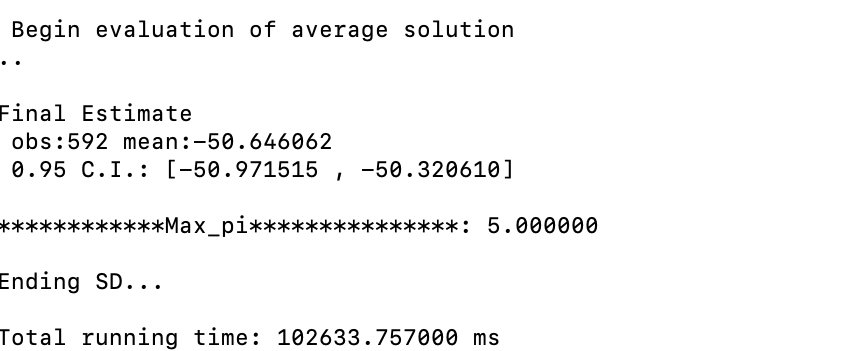


From the result, the SAA method generates the same output for x1 and x2. 9.92 for x1 and 49.6 for x2. The objective value is not the same, which is 50269.412. Running time is 0.07s.

**Stochastic Decomposition (SD)**

We write a pysp model and generate SMPS files and use sd solver to run. First we set the SD solver to run 1 replication. Next, we set the SD solver to run 30 replications.





From the result by SD solver, the objective value is 50557.273,running time is 2.150 second when the SD solver runs 1 replication.

The objective value is 50646, running time is 102.6 second when the SD solver runs 30 replications.

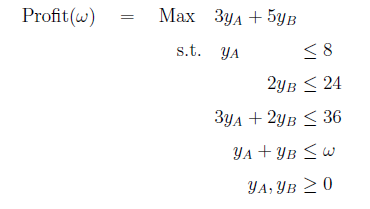
**5-fold Cross Validation**

After getting the objective value suggested by 3 models, our group takes a 5-fold cross validation. First we use the following formula to calculate variable wi for the validation set:

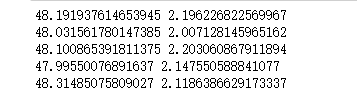


,, are coefficients generated by multiple linear regression. X1 and x2 in this formula is the objective x1 and x2 generated by LP and SAA.  is the error for the validation set.

Then we put variable wi to the following linear program to calculate the profit



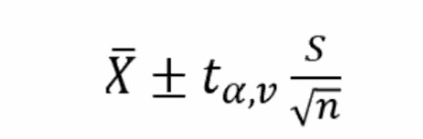
After we get the corresponding profit of 93 data points in the validation set (since we remove 7 outliers), we decide to take 5-fold cross validation for these profits. One thing important that we do not randomly drop 19 data points because randomization may cause more error. We split the 93 data points into 5 groups, we drop first 19 data, calculate the mean value and standard deviation of the rest 74 data points; next, we drop the second 19 data, then calculate the mean and standard deviation of the rest 74 data; then we drop the third group of 19 data, calculate the mean and standard deviation of the rest 74 data; we drop the fourth group of 19 data, calculate the mean and standard deviation of rest 74 data, then we drop last 17 data and calculate the mean and standard deviation of rest. We do it five times. Thus, we have mean value and five standard deviations. They are shown below .



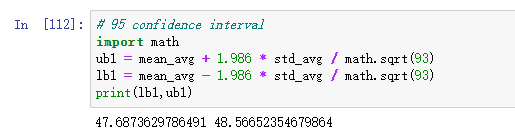
Then we get the mean value of the 5 mean objective profit and 5 standard deviations.

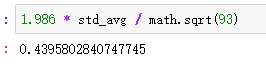


Then we use the formula to get 95 CI. Since the data set scale is not very large, we use t-distribution to find 95 Confidence Interval. Based on the T-value table, t value is 1.984.



From the formula above, we get the confidence interval is





From the result shown above, we draw a to form to show our results more clearly.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Methods | x1 | x2 | MPO | MVSAE | Time |
| MLR+ LP | 9.92 | 49.6 | 50441 | 48126±439.6 | 0.04s |
| MLR +SAA | 9.92 | 49.6 | 50269 | 48126±439.6 | 0.08s |
| MLR +SD | 9.92 | 49.6 | 50646 | 50321-50972 | 102s |

Due to the table of results, we find that the objective value provided by the SD method is the largest, and it falls in the 95 % confidence interval. The objective value provided by LP and SAA do not fall in the range of 95% confidence Interval. The Time for running LP is 0.04s, and the time to run SAA method is 0.08s, which is longer. That is reasonable because the Sample Average Method takes error into consideration. The time to run SD is 2s if it runs 1 replication, and 102 second if it runs the whole 30 replications.

**Lesson Learnt and Future Work**

Through this project, we learn the process to solve a real world problem by setting up the model, transforming to a linear model, get data, fit multiple linear regression to get coefficients, then use coefficients to solve linear programming. Also, we learn that before using coefficients, a validation is encouraged to do. Furthermore, we capture uncertain events and use that in the SAA method and SD method. After solving LP, we make a 5-fold cross validation, and split each fold successively to avoid errors caused by randomization.

There is also future work, we can try other algorithms besides linear regression to fit models of TV, Radio and Sales. We can take other methods of cross validation. Also based on the data, we can treat newspapers as predictors and change the constraints to change our model.