ISE 533 Notes: Integrative Analytics with Cross-Sectional Data

Suvrajeet Sen, Jiajun Xu, Yihang Zhang
Epstein Department of ISE
University of Southern California,
Los Angeles, CA 90089

Fall 2022

1 Scope

In the previous lecture, our discussion focused on the need for Integrated Analytics. In this set of notes, we will explore a concrete paradigm to facilitate an understanding of what may be necessary to bring data and decisions on a common footing. This essentially calls for statistical and optimization models to work under similar assumptions. If such synergy can be identified, then these modeling paradigms can be used to great advantage. These notes provide the basis for such integration. The "how" in these notes should be interpreted as "how to conduct" integrative analytics for situations with cross-sectional data. While we will present the assumptions underlying the models, we will not dwell on the mathematical processes which justify the methods discussed here.

2 The Advertising-Operations Integrated (AOI) Model

[1]

To illustrate the possibilities, let us return to the WYNDOR Glass example from the previous lecture, and consider the possibility that the company will be unable to sell however many doors it is able to produce. Instead, let us suppose that WYNDOR Glass can sell up to the number

[1] We would like to thank Yunxiao Deng for her contribution in formulating this model.

of doors demanded by its customers - a somewhat realistic circumstance. Since the demand may be guided by the amount of advertising, the total advertising budget may play an important role in the profitability of the firm. Suppose that management is considering an advertising budget of $200,000. The question then becomes how much advertising should be devoted to each type of outlet (TV/Radio) so as to maximize expected profit[2]. In order to undertake Integrative Analytics, we will first divide our data set into two halves: one for the purposes of training, and the other to conduct "out-of-sample tests" or validation. The validation step is intended as a "proxy" (or "stand-in") for predicting performance in the future.

Let x_1 and x_2 denote the TV and Radio expenditures respectively, and suppose that the total advertising budget is denoted $b = 200$ in thousands of dollars. Because of the need to have presence in both TV and Radio markets, we also have a constraint requiring that the Radio advertisement expenditures be at least some positive fraction of expenditures in TV ads, and this fraction is $\alpha \in (0, 1)$. Finally, we have lower and upper limits $[L_1, U_1]$ and $[L_2, U_2]$ which are included so as to allow the model to only include predictions within the range of expenditures for which we have data in the Advertising data set. In the training part of the data set, the ranges $[L_1, U_1] = [0.7, 296.4]$ and $L_2, U_2] = [0, 49.6]$. Assuming that the advertising costs are $c_1 = \$100$, $c_2 = \$500$ and $\alpha = 0.5$ an integrated budget allocation model may be written as follows.

$$\text{Max} \quad -0.1x_1 - 0.5x_2 + \mathbb{E}[\text{Profit}(\tilde{\omega})] \tag{1a}$$

$$\text{s.t.} \quad x_1 + x_2 \leq 200 \tag{1b}$$

$$x_1 - 0.5x_2 \geq 0 \tag{1c}$$

$$L_1 \leq x_1 \leq U_1, L_2 \leq x_2 \leq U_2 \tag{1d}$$

$$\tag{1e}$$

where $\tilde{\omega} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \tilde{\varepsilon}$ denotes the total sales, and this quantity appears in an operations model (see the model below) as a constraint on the amount of total production. In fact, one should treat the coefficients ($\{\beta_j\}$) of regression as random variables because there are estimation errors associated with these coefficients. While this can be accommodated within a decision/optimization framework, the standard errors of the coefficients $\{\beta_j\}$ are often small relative to the standard error associated with ε. This is particularly so for data sets containing a fairly large number of observations. We will return to this issue subsequently, but suffice it to say at this point that the

[2]We will discuss other risk management objectives subsequently.

profit depends on the choices x as well as the uncertainty captured via the random variable $\tilde{\varepsilon}$ (or equivalently, as a function of $\tilde{\omega}$). To see how this plays out in the operations stage of the model, consider the following slightly revised version of the WYNDOR Glass model (including a demand constraint) as shown in (2).

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

Those familiar with the original WYNDOR Glass production model in Hillier and Lieberman will recognize that the changes to the formulation appear in two forms: i) we have included a demand constraint (2d), and ii) we have scaled the right hand side coefficients of the H-L instance by multiplying the capacity coefficients by 2. The purpose of this re-scaling is merely to bring the two data sets (advertising and operations) in the same "ballpark". In any event, the interdependence between advertising and operations leads to two interconnected models (1 - 2), and one of these cannot be solved without solving the other.

3 Regression + Linear Programming: Can it Work?

For most optimization models in the literature (and most optimization courses) one makes the deterministic assumption that all data is certain. Because of this assumption, it is natural to first consider the case in which the random variable $\tilde{\omega}$ is replaced by its expectation. In this case, we would use the average sales as an input into a linear programming model, and with this assumption, would have the following LP as a planning model.

$$\mathbb{E}[\tilde{\omega}] = \mathbb{E}[\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \tilde{\varepsilon}] = \beta_0 + \beta_1 x_1 + \beta_2 x_2. \tag{3}$$

Note that the random variable disappears from the above equation because in MLR we assume that $\mathbb{E}[\varepsilon] = 0$. Given that the right hand side above is a deterministic linear function, one could use it as part of an LP model as shown below.

$$\begin{aligned}
\text{Max} \ & -0.1x_1 - 0.5x_2 + \ 3y_A + 5y_B && && (4a)\\
\text{s.t.} \ & x_1 + x_2 && \leq 200 && (4b)\\
& x_1 - 0.5x_2 && \geq 0 && (4c)\\
& y_A && \leq 8 && (4d)\\
& 2y_B && \leq 24 && (4e)\\
& 3y_A + 2y_B && \leq 36 && (4f)\\
& -\beta_1 x_1 - \beta_2 x_2 + y_A + y_B && \leq \beta_0 && (4g)\\
& y_A, y_B \geq 0 && && (4h)\\
& L_1 \leq x_1 \leq U_1, \quad L_2 \leq x_2 \leq U_2 && && (4i)
\end{aligned}$$

The linear regression coefficients we obtained are $\beta_0 = 3.1470, \beta_1 = 0.046, \beta_2 = 0.184$.

Before using these coefficients in an optimization model, it is prudent to put in place a validation mechanism so that we will know how we plan to estimate the effectiveness of any allocation of the advertising budget $(\hat{x}_1, \hat{x}_2)$. Then the following is our completely data driven validation approach.

1. For each data point $(x_{1i}, x_{2i}, \omega_i)$ in the validation set, calculate

$$\varepsilon_{vi} := \omega_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i} \tag{5}$$

2. For each i in the validation set, calculate

$$\omega_i := \beta_0 + \beta_1 \hat{x}_1 + \beta_2 \hat{x}_2 + \varepsilon_{vi} \tag{6}$$

3. For each data point $(x_{1i}, x_{2i}, \omega_i)$ in the training set, calculate

$$\varepsilon_{ti} := \omega_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i} \tag{7}$$

   (Note that the above equation assumes that the random variable ε_{vi} and ε_{ti} do not change with different values of x. This is also part of the assumptions of MLR.)

4. Using the samples ε_{vi} and ε_{ti}, perform the F-test to verify whether the errors are drawn

from the same distribution.

5. If the null hypothesis (that the errors are from the same distribution) is rejected, then, we should use a larger proportion of the data set for training, and repeat the process as suggested in the overall Integrative Analytics framework. Else, we illustrate the fit using a Q-Q plot and continue.

6. For each ω_i calculated in (6) solve the LP associated with Profit(ω_i) in (2). Using this data stream calculate the average objective function value $c_1\hat{x}_1 + c_2\hat{x}_2 + \frac{1}{N}\text{Profit}(\omega_i)$, as well as the standard deviation.

7. Report the Data Driven Average and 95% Model Predicted Objective (MPO). [3]

Then the allocation of budgets to TV and Radio advertising suggested by the LP model is shown in Table 1 [4].

x_1	x_2	MPO (in $)	MVSAE (in $)
173.48	26.52	$41,391.3	$[38,810, 40,909]

Table 1: Deterministic Solution using Regression + Linear Programming

Several questions arise from the output reported in Table 1:
a) Is the predicted objective value ($41,391.3) reported in Table 1 the maximum average profit? or b) Is the predicted objective value ($41,391.3) the maximum profit among all possible average sales values? A little reflection reveals that it is b). To see this, note that (4g) requires $y_A + y_B \leq \beta_0 + \beta_1 x_1 + \beta_2 x_2$. This implies that production should be less than or equal to the average sales for any vector (x_1, x_2). Hence the deterministic LP chooses that vector of advertising expenditures for which the average sales yields the highest profit.
The next question is "How does the reported/predicted (objective) value from LP (4) compare with the average profit associated with the proposed plan of spending $173,480 on TV ads, and $26,520 on Radio ads? As shown in Table 1, the average profit associated with the above-mentioned advertising expenditures is approximately $39,000. In other words, the value reported by the deterministic LP ($41,391.3) is higher than the average profit one expects if such a plan is adopted.

[3] MPO is the objective function estimate provided by the optimization model. In case of LP this will be a point estimate. If SAA is run using one sample, then, again the MPO will be a point estimate. Since SD is run with replication, it will report the its MPO as a 95% CI.
Model Validation Sample Average Estimate (MVSAE): This is a 95% CI of the Sample Average of the SLP using errors from the validation set. Since this should be calculated in the SD code through an evaluation run, the calculation would follow the standard equation (see (19)) in my paper with Yifan (attached).

[4] Depending on the choice of training validation split answers may be slightly different

That is, the prediction from the deterministic LP is rosier than what could be realistically expected in an uncertain environment. For a maximization problem as in the above case, the deterministic LP yields a higher prediction than the average profit one should expect in an uncertain environment. This observation is a manifestation of a theorem known as Jensen's inequality. In order to gain a fuller understanding of these differences, let us examine some assumptions underlying MLR.

4 Assumptions underlying Multiple Linear Regression

Clearly a deterministic LP fall short because of the inconsistencies mentioned above. Since the interface between (1 - 2) is via ω (or more accurately the random variable $\tilde{\omega}$), it is important for us to reconcile the assumptions underlying the paradigms of statistical estimation and optimization.

1. Linearity: As the name implies, a linear fit must be deemed appropriate for any MLR model.

2. Independence: The error terms in a regression model are assumed to be independent random variables.

3. Normality: Another standard assumption in regression is that the error (i.e., difference between the "least-squares" estimate, and the observed data) is approximately normally distributed. From a visual perspective, one can generate a quantile-v-quantile (Q-Q) plot which compares how the quantiles of the empirical (data) distribution compares with quantiles of a normal distribution. If the resulting plot is close to a straight line, then one can justify the normality assumption. One can also quantify the degree of closeness of the empirical distribution to normality by using a Kolmogorov-Smirnoff (K-S) test which measures the distance of the empirical distribution from the normal (with mean 0, and variance σ^2). Of course, the quantity σ is usually unknown, and one estimates it from the data as discussed below.

4. Error Distribution: This refers to the property that ε, the error distribution does not vary with the choice of x.

If all these assumptions hold, it can be shown that the variance $\sigma^2 = \mathbb{E}[\frac{\sum_i \varepsilon_i^2}{N-n-1}]$. Hence with data points indexed by i, one can estimates σ^2 as follows.

$$s^2 = \frac{\sum_i^N \varepsilon_i^2}{N-n-1}, \tag{8}$$

where, N is the number of data points, and $n+1$ is the total number of parameters $\{\beta_j\}_{j=0^n}$ in MLR.

Ordinarily, for somewhat large data sets, one should be able to justify that the parameters yield stable estimates of the estimated mean $\beta_0 + \sum_j \beta_j x_j$, and the random variable $\tilde{\varepsilon}$ can be approximated by a normal distribution when the number of degrees of freedom is large enough. In the statistical literature, it is customary to suggest that when the degrees of freedom (i.e. $N-n-1$) exceeds about 60, the parameters $\{\beta_j\}_{j=0}^n$ are quite stable, and the errors ε are approximately $\mathbf{N}(0, s^2)$. Thus, for the example with the advertising data set, the number of data points $N = 100$ (for the training data set), and the number of parameters being used is $n+1 = 3$. Hence, the number of degrees of freedom far exceeds the suggested 60, and as a result, ε in such examples may be reasonably approximated by using normal random variates using $\mathbf{N}(0, s^2)$.

5 Regression + Stochastic Linear Programming: Sample Average Approximation

In order to integrate MLR with a decision model, it is clear that an optimization approach should accommodate the assumptions of the MLR framework. As shown earlier, a deterministic optimization framework is at odds with MLR. What about a stochastic framework? Indeed, the model stated in (1-2) can be formally classified as a *Stochastic Linear Programming (SLP)* problem.

As the name suggests, SLP is an outgrowth of deterministic linear programming, and is intended to allow a representation of data uncertainty using discrete random variables in an optimization model. The traditional form of such SLP, as stated in a 1955 article by George Dantzig[5] was intended to be applicable to optimization problems with finitely many outcomes because the computational approaches of the time were only based on the Simplex Method, an algorithm that requires finite dimensions. However, by using the empirical distribution to approximate the distribution of any random variable, it has been shown that finite dimensional SLPs are able to provide asymptotically consistent decisions and objective function estimates for planning models based on regression (as in the case of the AOI model). However, in some instance, the sample size associated with empirical distributions can be extremely large, and in such cases, using LP solvers to solve SLP models can become treacherous. We shall return to address this situation in a bit.

[5]Dantzig is the father of the Simplex method, and many more advances in LP

For the moment, let us consider the traditional case of SLP with finitely many outcomes. In this case, the AIO model in (1) should replace the expectation operator ($\mathbb{E}$) by a finite representation of the average objective, that is, $\sum_{i=1}^{N} p_i \text{Profit}(\omega_i)$, where $0 \leq p_i \leq 1$ denotes the probability of outcome i, and of course, $\sum_{i=1}^{N} p_i = 1$.

Returning now to the MLR setting, we make two important observations regarding the use of SLP. i) First and foremost, we recognize that in decision process for AIO as well as other SLP models, one is interested in two kinds of decisions: those which must be made before the uncertainty clears (e.g. before actual demand/sales numbers are revealed), and those decisions that are made in response to the observed outcomes (e.g. production plan to meet sales), ii) in the context of statistical applications such as MLR, the probability estimates p_i are usually replaced by the empirical distribution of demand/sales, that is, $p_i = \frac{1}{N}$, where N denotes the number of data points. Without MLR, the probabilities in SLP are difficult to predict. Assuming that an MLR model has already produced regression coefficients $\{\beta_j\}_{j=0}^{n}$, an appropriate data driven representation of errors is given by

$$\omega_i := \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon_{ti} \tag{9}$$

where ϵ_{ti} are given in (7). Note that (9) assumes that most of the uncertainty in the MLR model is due to ε, although uncertainties can also appear in other coefficients as well. However, as the data set increases in size, it is the standard error of ε which tends to dominate uncertainty, and for such applications it is appropriate to use SLP with (9), capturing the uncertain events via ε, with its empirical distribution $\frac{1}{N}$ replacing the probabilities. A statistical approximation of the SLP model, which should classified as a *sample average approximation - SLP/SAA*, can then be stated as follows.

$$\text{Max} \ -0.1x_1 - 0.5x_2 + \frac{1}{N}\sum_{i=1}^{N} \text{Profit}(\omega_i) \tag{10a}$$

$$\text{s.t.} \quad x_1 + x_2 \leq 200 \tag{10b}$$

$$x_1 - 0.5x_2 \geq 0 \tag{10c}$$

$$L_1 \leq x_1 \leq U_1, L_2 \leq x_2 \leq U_2 \tag{10d}$$

where the bounds $[L_1, U_1]$ and $[L_2, U_2]$ are once again derived from the advertising data set, and must be the same as in (1). Moreover, each outcome of the "Profit" random variable is defined as

$$
\begin{aligned}
\text{Profit}(\omega_i) \quad = \quad \text{Max} \quad & 3y_A + 5y_B && \text{(11a)}\\
\text{s.t.} \quad & y_A \le 8 && \text{(11b)}\\
& 2y_B \le 24 && \text{(11c)}\\
& 3y_A + 2y_B \le 36 && \text{(11d)}\\
& y_A + y_B \le \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon_{ti}, \forall i && \text{(11e)}\\
& y_A, y_B \ge 0 && \text{(11f)}
\end{aligned}
$$

It is important to note that unlike the model in (4), every data point i gets to choose its own production level (y_{Ai}, y_{Bi}) because $\text{Profit}(\omega_i)$ is obtained by solving the LP in (11) for each i. Since we use the maximum profits for each ω_i, it follows that the SLP in (10) the Sample Average Approximation (SAA) provides a statistical estimate of Expected Profit.

At this point it is appropriate to ask the question posed as the title of this lecture: "How" does one solve the integrated model (10 - 11)? For the smallest of instances (such as our AOI instance), one can write the formulation in (10 - 11) as a large scale linear program as follows.

$$
\begin{aligned}
\text{Max} \quad & -0.1x_1 - 0.5x_2 + \frac{1}{N}\sum_{i=1}^{N}(3y_{Ai} + 5y_{Bi}) && \text{(12a)}\\
\text{s.t.} \quad & x_1 + x_2 \le 200 && \text{(12b)}\\
& x_1 - 0.5x_2 \ge 0 && \text{(12c)}\\
& y_{Ai} \le 8 \quad i = 1, \ldots, N && \text{(12d)}\\
& 2y_{Bi} \le 24 \quad i = 1, \ldots, N && \text{(12e)}\\
& 3y_{Ai} + 2y_{Bi} \le 36 \quad i = 1, \ldots, N && \text{(12f)}\\
& -\beta_1 x_1 - \beta_2 x_2 + y_{Ai} + y_{Bi} \le \beta_0 + \varepsilon_{ti} \quad i = 1, \ldots, N && \text{(12g)}\\
& L_1 \le x_1 \le U_1, \quad L_2 \le x_2 \le U_2, y_{Ai}, y_{Bi} \ge 0. && \text{(12h)}
\end{aligned}
$$

For the AOI data set, the above SLP/SAA is very manageable. Thus if $N = 100$ (as in the AOI training set), we have 202 variables (i.e. x_1, x_2, and 100 pairs of (y_{Ai}, y_{Bi})), each with their bounding constraints, and another 200 constraints (associated with a capacity and a demand constraint) for each data point i. Such an instance is well within the realm of ordinary LP software because

the number of variables and constraints are relatively small (in the hundreds).

After solving SAA model we get the decisionsx_1 and x_2 as well as the predicted Objective value. We repeat the validation steps to get the MVSAE.

Methodology	x_1	x_2	MPO (in \$)	MVSAE
Deterministic LP	173.48	26.52	\$41,391.3	\$[38,810, 40,909]
SLP with SAA	191.027	8.793	\$ 40,179	\$[41,375, 43,362]

Table 2: SLP Solution using SAA

The plan obtained from the SLP/SAA is shown in the second row of Table 2. It is important to notice that the differences in recommendations from the two alternative formulations: the deterministic LP model and the SLP/SAA approach. Firstly, the decisions (x_1, x_2) are themselves quite different. Moreover, the validated expected profit using the plan from the deterministic LP is about \$38,000, whereas, the validated expected profit from the SLP/SAA model is significantly higher (by about \$2,179). This amounts to a 5% difference on average.

6 Regression + Stochastic Decomposition: The Case for Replications

In this section, we will introduce another feature which is common to SLP, and other stochastic optimization settings, such as simulation optimization. The feature we wish to highlight here is the need for replications. In most formal expositions of SAA, the analyst is encouraged to undertake replications of the SAA process because there is a certain degree of arbitrariness with the way in which the data set is managed. For instance, dividing the data set into two so that one can use one half for training and another for testing could be looked upon as somewhat arbitrary. Even if one used some other fraction (say 80% training and 20% validation) may still raise similar questions. In statistical analysis, one way to reduce variance is by using replications; in other words, one can obtain low variance estimates of the sample mean (of the optimum value) by replicating the optimization process using randomization. However, for reasons discussed in subsequent lectures, replications can be computationally demanding, and in the absence of high-end computing/super computers, most applications often report the results of only one training run. One method which automatically undertakes replications in SLP is called Stochastic Decomposition (SD). There are many other features of SD which we will discuss at various points in these

lectures; for this section, we will simply focus on some basics: assumptions of SD, how to run SD, and finally, we will present a comparison of results from SD with the outputs presented earlier.

6.1 Assumptions for Stochastic Decomposition

As suggested earlier, using standard LP software on practical SLP problems can become a daunting task. For this reason, most SLP models are solved using decomposition: either deterministic or stochastic. The former are usually sufficient for instances with a few outcomes/data points, whereas, the size of the data set is not a concern with the latter; i.e., arbitrarily large data sets are allowed in SD. The assumptions below are stated in the context of SD as a solver for SLP models in which randomness appears only through the MLR model, and nothing else.

- Boundedness: In the setting of MLR, suppose that we first divide the data sets into two halves. Then, $N = 100$, and one can scan the data set for $\ell_j = \min\{x_{i,j}, i = 1, \ldots, N\}$, and $u_j = \max\{x_{i,j}, i = 1, \ldots, N\}$. Then one may define $X = \{x \mid \ell_j \leq x_j \leq u_j, j = 1, \ldots, n\}$, In the advertising example, recall that the number of training data points is $N = 100$, and the number of predictors is $n = 2$. So, the space of decisions X is in $\Re^2$, and the upper and lower bound for each predictor (TV and Radio) is obtained using the min and max values in the data set.

- Fixed Recourse: First a word about terminology. The LP in the second stage (e.g.,(11)) is referred to as the Recourse Problem. A model is said to possess the Fixed Recourse property when all the random elements of the recourse problem appear only on the right hand side; see (11) for example. A quick glance at (11) reveals that randomness only appears on the right hand side (even if $\beta_0, \beta_1, \beta_2$ are all random). Hence this assumption is satisfied.

- Relatively Complete Recourse: Here it is assumed that the recourse problem (e.g.,(11)) remains feasible for all values of (x, ε). For instance in (11), $\beta_1 x_1 + \beta_2 x_2$ is non-negative for all $x \in X$, and $\beta_0 + \varepsilon_i$ are non-negative for all i, the constraints in the second stage corresponding to the regression equation is $y_A + y_B \leq \delta$, where $\delta \geq 0$. Hence (11) is feasible for possible choices of (x, ε).

- Optimistic Lower Bound: In many minimization problems, the second stage cost is often bounded below by 0 in the most optimistic case. For a "Max" problem, optimism requires that we specify an upper bound in the best case scenario. For the WYNDOR Glass data,

which is Profit Maximization model, an optimistic profit could be the result of selling 30,000 units (which is higher than any sales amount reported in the data). Moreover, suppose that one makes a profit of $5 per unit (which is the highest per unit profit in the data set). Then for the WYNDOR Glass data, the optimistic bound to use would be a profit of $150 (thousand).

The rest of the assumptions of SD are inherited from the SLP assumptions discussed in section 5.

6.2 Running SD

It is not difficult to see that the size of the LP given in (12) depends on the types of different doors (the size of the vector y), and the number of points (N) in the data set. For textbook instances such as the WYNDOR Glass example, standard LP software suffices. However, with larger data sets (e.g., firms like Proctor and Gamble (P&G) and Kroger) which produce/order many types/sizes of products, and have very large data sets for each of product, the size of the stochastic LP (12) can become arbitrarily large. In order to setup such large scale models, and manage the data, it is difficult to use standard LP software (e.g. AMPL, CPLEX etc.) to solve SLP.

SD is a sequential sampling algorithm based on sampling from the training set, drawing data points by sampling from the empirical distribution. As the algorithm builds its objective (value) function approximation, it uses a variety of non-parametric optimality tests before concluding any sampled run. It then repeats this process as many times as desired by the user. We recommend 30 re-sampling runs, unless the first few produce very little variability in the solution. There are two solutions typically reported at the end of these runs: one is called the compromise solution, and the other is called the mean solution. When the two are very similar, both solutions can be expected to be very good. Such a solution is refereed to as the Compromise Decision.

6.3 Results

The results for all three approaches, the Deterministic LP, SLP with SAA and SLP with SD are summarized in Table 4, although the first two rows are repeated here for the sake of convenience. The first observation we make is that SLP models, and almost any other realistic stochastic optimization model, are difficult to validate. However, one should expect that the model provide some consistency between its predicted and the validated objective values. Note that both the deterministic LP, as well as the SLP with SAA do not predict the objective function accurately enough to fall within their respective 95% CI (which we should include in Table 3). However, we

find that the prediction from SD is much closer to the Validated Objective function estimates. The most likely reason for this difference is that replications from SD produce a decision referred to as the compromise decision which is known to have better variance reduction properties than an optimal solution from any one run.

Methodology	x_1	x_2	MPO (in \$)	MVSAE	Time(s)
Deterministic LP	173.48	26.52	\$41,391.3	\$[38,810, 40,909]	0.211
SLP with SAA	191.027	8.793	\$ 40,179	\$[41,375, 43,362]	0.497
SLP with SD	191.4	8.6	\$ 42,045 \$ [41,110, 42,980]	\$[41,384, 43,276]	24.083

Table 3: Comparison Solutions from Deterministic LP, SLP using SAA, SLP using SD

	x_1	x_2	MPO	MVSAE	0.95C.I. of Sales	Obj. Interval
LP	173.48	26.52	\$ 41,391	\$[38,810, 40,909]	[15658.7, 16310.7]	\$[40,368, 41,392]
SAA	191.027	8.793	\$ 40,179	\$[41,375, 43,362]	[13021.4, 14029.2]	\$[39,565, 45,488]
SD	191.4	8.6	\$ 42,045	\$[41,384, 43,276]	[12999.4, 14014.4]	\$[39,558, 42,603]

Table 4: Obj. Interval from Deterministic LP, SLP using SAA, SLP using SD

7 Putting it all together: Data, Decisions and Validation

Let us first summarize what we have accomplished in this lecture.

- Paradigms like Regression and Stochastic Programming are founded on very similar assumptions, and should provide a mathematically strong foundation for Integrative Analytics
- SLP/SAA provides a more defensible approach to integration with Regression than Deterministic Linear Programming
- The variance reduction features within SLP/SD are able to generate decisions which are more likely to produce lower variability in the validation phase.

Here is a flowchart that describes the workflow for this project.

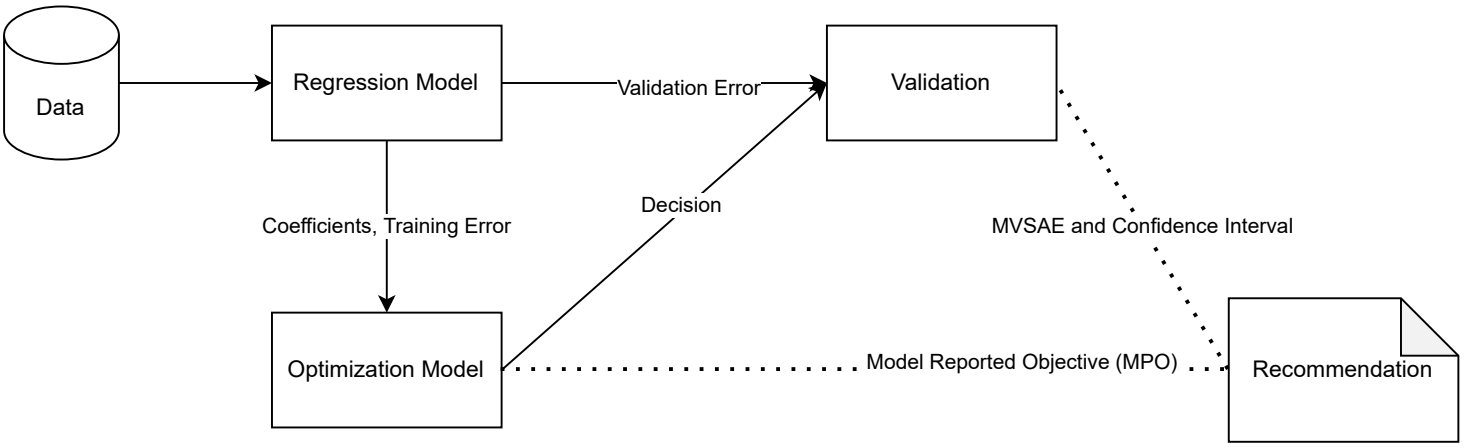

Figure 1: Workflow

7.1 Step-by-step Details

7.1.1 Deterministic Linear Programming

- Descriptive Analytics:
  1. Split data into training and validation set. A 50-50 split is recommended since we need enough data to estimate the validation objective later.
  2. Use the `lm` program in `R` (or your favorite package in Python/Julia) to build a linear regression model.
  On the training set, use TV and radio as regressors to obtain the linear regression coefficients.
  3. Calculate ε_{ti} and ε_{vi} by equation (5) and (7).

- Prescriptive Analytics: Use deterministic linear programming to solve the allocation model
  1. Write down the deterministic LP Model (4) in Julia.
  2. Solve the LP. Report the objective as MPO.

- Validation:
  1. Do F-test of ε_{ti} and ε_{vi} to check whether they are drawn from the same distribution. Alternatively, plot the histogram of ε_{ti} and ε_{vi} .
  2. Plot the Q-Q graph of ε_{ti} and ε_{vi} to check whether they are normally distributed. (Are there any outliers?)

3. Using the ε_{vi} from Descriptive Analytics, calculate ω_i as in (6). Calculate $Profit(\omega_i)$ in each scenario with $\hat{x}_1, \hat{x}_2$ **fixed** by solving second stage LPs. Calculate $-0.1\hat{x}_1 - 0.5\hat{x}_2 + \frac{1}{N}\sum_{i=1}^{N} Profit(\omega_i)$. This gives the Model Validation Sample Average Estimate (MVSAE) and a confidence interval of the profit.

7.1.2 SAA

- Descriptive Analytics: Same steps as the previous method.

- Prescriptive Analytics:
  1. Write down the all-in-one LP with the training errors ε_{ti} as the uncertainty. See (12).
  2. Solve the LP. Report MPO.

- Validation:
  Repeat the validation steps in Deterministic Linear Programming, but this time fix $\hat{x}$ to the solution you obtained from the all-in-one large LP.

7.1.3 SD

- Descriptive Analytics
  1. Split the data into training and validation set.
  2. Use the training segment to identify a Linear Regression fit. Calculate ε_{ti}.

- Prescriptive Analytics

  1. Create a direct_model(CPLEX.Optimizer()) in Julia and import all the constraints and objective specified in (4) as a template.
  2. Create a function that returns a random sample of ε_{ti} each time it is called. The return type should be a OneRealization object, and should bound the RHS of constraint

(4g) to $\beta_0 + \varepsilon_{ti}$. Specify the mean of the RHS, which is β_0.

3. Specify which Position (constraint and variable) in the template separates the first and second stage. In this case, should be the constraint (4d) and the variable y_A.

4. (Optional) In the TwoSD folder, open twosd/config.sd, and change MULTIPLE_REP to 1 to have SD solver calculate compromise decisions using multiple replications.

5. Run SD to obtain the decision. The decisions are stored in the current working directory. The MPO will be displayed on the screen.

- Validation:
  Repeat Validation steps.