Optimal allocation in regression models with cost consideration

Yujing Du
College of Electrical and Information Engineering, Jilin Agricultural Science and Technology University, Jilin, China, 132101
Email: duyj219@163.com

Abstract. In this paper, we discuss the optimal allocation problem in a multi-level experiment when the cost of running the experiment at each level is taken into consideration. An objective function with cost consideration is proposed and the properties of the objective function are discussed. The procedure to the algorithm to obtain the optimal allocation is presented. Numerical examples are used to illustrate the methodology developed here.

1. Introduction
The optimal allocation problem in a multi-level experiment is closely related to the optimal design problem for a general regression model. For regression model, researchers considered different methods to handle various cost structures in specific experimental design settings. For instance, the normalization of the variance-covariance matrix for each cohort in clinical designs is proposed [1-2], the optimal designs with a restriction on the total cost of the experiment is discussed [3-4], the work is focused on the optimal design by considering the cost of the study based on a linear regression models with random effects, and the optimal criteria is considered through maximizing the information under the restriction on the total cost, or minimizing the cost function under the specified precision [5], however, they did not consider the balance between the cost and the information measure simultaneously. Moreover, simple computational algorithm to obtain the optimal design was not provided explicitly.

In this paper, we include a balance factor in the objective function to balance between the precision in parameter estimation and the cost of the experiment when linear regression model is used for statistical analysis. We also discuss the numerical computational algorithm under the proposed objective function. For recent development on numerical computational algorithm for optimal allocation problem in linear regression model, the monotonic convergence for a general class of computational algorithm for D-optimality design is discussed, but the algorithm suffers from the slow convergence problem [6]. Recently, D-optimality and A-optimality criteria are used to propose a computational algorithm and the related mathematical properties are discussed [7]. The main focus of this paper is to obtain the optimal allocation for a multi-experiment with cost consideration.

2. Models and objective function with cost consideration
Consider an experiment with response variable \( Y \) and \( p \) explanatory variables \( \mathbf{x} = (x_1, x_2, \ldots, x_k) \), the relationship between \( Y \) and \( \mathbf{x} \) can be modelled by a linear regression model:

\[
Y = \mathbf{x}\beta + \epsilon
\]
Where $\beta \in \mathbb{R}^p$ is a $p$-dimensional vector of the regression parameter, and $\varepsilon$ is the random error with $E(\varepsilon) = 0$ and $\text{Var}(\varepsilon) = \sigma^2$. Without loss of generality, we can rewrite the combinations of different levels of the explanatory variables into $k$ experimental conditions (or design points) denote by $x_l = (x_{l1}, x_{l2}, ..., x_{lp})$, $l = 1, 2, ..., k$, where $x_{il}$ is the $l$-th levels of the $i$-th explanatory variable. If the intercept term is included in the regression model, then we are looking at $p$-1 explanatory variables and $x_{il}$, $l = 1, 2, ..., k$. Suppose $n_l$ items are assigned for testing at experimental condition $x_l$, $l = 1, 2, ..., k$, and the corresponding values of the response variables, say $y_l = (y_{l1}, y_{l2}, ..., y_{ln_l})'$ are observed, then

$$y_{lj} = x_{lj}' \beta + \varepsilon_{lj}, l = 1, 2, ..., k, j = 1, 2, ..., n_l.$$

The method of least squares can be used to estimate the vector of model parameter $\beta$. The least squares method estimate of the regression parameter $\hat{\beta}$ can be by

$$\text{argmin}_{\beta} \left\{ \sum_{l=1}^{k} \sum_{j=1}^{n_l} (y_{lj} - x_{lj}' \beta)^2 \right\},$$

which can be expressed as

$$\hat{\beta} = \left( \sum_{l=1}^{k} n_l x_l' x_l \right)^{-1} \left( \sum_{l=1}^{k} \sum_{j=1}^{n_l} y_{lj} x_l \right).$$

The variance-covariance of the estimate of $\beta$ can be expressed as

$$\sigma(\hat{\beta}) = \sigma^2 \left( \sum_{l=1}^{k} n_l x_l' x_l \right)^{-1}.$$

In experimental planning, suppose we have $N$ items available and the flexibility in the assigning these items to the $k$ experimental conditions, i.e. choosing $(n_1, n_2, ..., n_k)$ with $\sum_{l=1}^{k} n_l = N$. The optimal allocation problem is to find $(n_1, n_2, ..., n_k)$ that minimizes an objective function related to the variance-covariance matrix $\sigma(\beta)$. The optimal allocation problem has been studied extensively in the literature for different objective functions with fixed value of $N$ such as the determinant of the Fisher information matrix (i.e. D-optimal) and the trace of the variance-covariance matrix (i.e. A-optimal).

In this study, we consider that the total number of items in the experiment $N = \sum_{l=1}^{k} n_l$ is not fixed, and consider a practical situation that the experimental cost at different experimental conditions are different. Note that cost of the experiment can be the cost of running the experiment or the risk involved in the experiment. For instance, if we are testing the responses of patients to a drug with $k$ different dose levels, the higher the dosage of the drug used, the higher the risk of a patient getting side-effect and/or the higher the cost to create the drug. Therefore, in addition to the variance-covariance matrix $\sigma(\beta)$, the cost of the experiments at different experiment levels should be incorporated into the objective function. In order to take the cost of the experiments at different experiment at different level into account, we propose a general objective function with the following form:

$$\log | \left( \sum_{l=1}^{k} n_l x_l' x_l \right)^{-1} | + \rho \sum_{l=1}^{k} r_l n_l$$

(1)

Where $\rho \geq 0$ is a factor that balance between the cost and the variation in estimation; $r_l \geq 0$ is the cost of running the experiment for one item on experimental level $l, l = 1, 2, ..., k$; and $n = (n_1, n_2, ..., n_k)$, usually $n_l \geq 0, l = 1, 2, ..., k$ are positive integer. Based on the objection function (1), we can simultaneously consider the accuracy in parameter estimation and the cost of the experiment.

The objection function is relatively simple and possesses some nice properties as presented in the following sections.
Instead of considering \( n_l \) as an integer, we consider the general case that \( w_l (l = 1,2, ... k) \) is a nonnegative real number with the objective function
\[
g(w) = \log \left( \sum_{l=1}^{k} w_l x_l x_l' \right)^{-1} + \rho \sum_{l=1}^{k} r_l w_l,
\]
hence, we are aiming to obtain the solution for \( \arg \min_w g(w) \).

### 3. Results and discussion

In this section, we present the results for optimal allocation that minimizes the objective function in equation (2).

**Theorem 1.** The vector \( w^* = (w_1^*, w_2^*, ..., w_k^*) \) is the solution of \( \arg \min_w g(w) \) if and only if
\[
\begin{cases}
x_l^*\left(\sum_{l=1}^{k} w_l^* x_l x_l'\right)^{-1} x_l - \rho r_l = 0 & \text{for } w_l^* \neq 0, \\
x_l^*\left(\sum_{l=1}^{k} w_l^* x_l x_l'\right)^{-1} x_l - \rho r_l \leq 0 & \text{for } w_l^* = 0.
\end{cases}
\]

**Proof:** Note that \( g(w) \) is a concave function. We can obtain the partial derivative of \( g(w) \) with respect to \( w_l \) is
\[
\frac{\partial g(w)}{\partial w_l} = -\text{trace}\left[\sum_{l=1}^{k} w_l x_l x_l'\right] + \rho r_l = -x_l^*\left(\sum_{l=1}^{k} w_l x_l x_l'\right)^{-1} x_l + \rho r_l,
\]
the second order partial derivative as
\[
\frac{\partial^2 g(w)}{\partial w_l^2} = x_l^*\left(\sum_{l=1}^{k} w_l x_l x_l'\right)^{-1} x_l x_l'\left(\sum_{l=1}^{k} w_l x_l x_l'\right)^{-1} x_l > 0.
\]
By the Kuhn-Tucker conditions [8], \( w^* = (w_1^*, w_2^*, ..., w_k^*) \) is the solution to minimize equation (2) if and only if
\[
\left(\frac{\partial g(w)}{\partial w_1} \frac{\partial g(w)}{\partial w_2} \cdots \frac{\partial g(w)}{\partial w_k}\right)'|_{w=w^*} (w - w^*) \geq 0, \text{ for } w_1 \geq 0, w_2 \geq 0, \ldots, w_k \geq 0.
\]
Since
\[
\left(\frac{\partial g(w)}{\partial w_1} \frac{\partial g(w)}{\partial w_2} \cdots \frac{\partial g(w)}{\partial w_k}\right)'|_{w=w^*} (w - w^*) = \sum_{l=1}^{k} \left[ -x_l^*\left(\sum_{l=1}^{k} w_l^* x_l x_l'\right)^{-1} x_l + \rho r_l \right] (w_l - w_l^*),
\]
The theorem follows.

**Remark 1:** Form Theorem 1, if \( w_1^* \) is the solution to \( \arg \min_w g(w) \) corresponding to \( r_{11}, r_{12}, ..., r_{1k}, \) \( w_2^* \) is the solution to \( \arg \min_w g(w) \) corresponding to \( r_{21}, r_{22}, ..., r_{2k}, \) and \( r_{2l} = cr_{1l} \) for \( l = 1, 2, ..., k \), where \( c \) is a constant, then \( cw_2^* = w_1^* \).

**Remark 2:** If \( r_1 = r_2 = \cdots = r_k = r^* \), and \( r \) is a known number, the result in the Theorem 1 is the same as the result presented in reference [7].

### 4. Simulation studies

In order to study the optimal results basing on the Theorem 1, an extensive simulation study is performed. We generate \((x_1, x_2, ..., x_k)\) from a normal distribution with number of design \( k=10,20,30 \)
and number of covariate \( p=3 \) with \( p<k \). The value of \( r \) is generated uniform distributed in between 0.001 and 1, i.e., \( U(0.001,1) \). The stopping criteria of the algorithm are set to:

\[
\max \left\{ w_{i}^{(h)} - w_{i}^{(h-1)} \right\} < 0.001.
\]

For each combination of \( p \) and \( k \), 50 replications are simulated. The number of iterations and the elapse time required to obtain optimal allocation are recorded. The average number of iterations and the average elapse time are presented in Table 1, Table 2 and Table 3 for \( p=3 \), \( p=5 \) and \( p=7 \), respectively.

**Table 1.** Simulated results of average number of iterations and elapsed time when \( p=3 \).

| k  | \( \rho \) | Average number of iterations | Average number of elapsed time |
|----|--------|-------------------------------|------------------------------|
| 10 | 0.1    | 85.6112                       | 0.1050                       |
|    | 0.2    | 74.6572                       | 0.0894                       |
|    | 0.3    | 65.9664                       | 0.0878                       |
|    | 0.5    | 56.5384                       | 0.0678                       |
| 20 | 0.1    | 129.1200                      | 0.3215                       |
|    | 0.2    | 107.9968                      | 0.2899                       |
|    | 0.3    | 96.2896                       | 0.2245                       |
|    | 0.5    | 83.1200                       | 0.2143                       |
| 30 | 0.1    | 151.6570                      | 0.5889                       |
|    | 0.2    | 133.0236                      | 0.4867                       |
|    | 0.3    | 117.2780                      | 0.3948                       |
|    | 0.5    | 101.0484                      | 0.3468                       |

**Table 2.** Simulated results of average number of iterations and elapsed time when \( p=5 \).

| k  | \( \rho \) | Average number of iterations | Average number of elapsed time |
|----|--------|-------------------------------|------------------------------|
| 10 | 0.1    | 80.1732                       | 0.0978                       |
|    | 0.2    | 65.7756                       | 0.0808                       |
|    | 0.3    | 60.4544                       | 0.0734                       |
|    | 0.5    | 51.8894                       | 0.0639                       |
| 20 | 0.1    | 127.4328                      | 0.3042                       |
|    | 0.2    | 105.7084                      | 0.2536                       |
|    | 0.3    | 93.9220                       | 0.2203                       |
|    | 0.5    | 82.9808                       | 0.2001                       |
| 30 | 0.1    | 147.0689                      | 0.5255                       |
|    | 0.2    | 125.6376                      | 0.4488                       |
|    | 0.3    | 109.5044                      | 0.3908                       |
|    | 0.5    | 95.4320                       | 0.3427                       |
Table 3. Simulated results of average number of iterations and elapsed time when p=7.

| k   | ρ  | Average number of iterations | Average number of elapsed time |
|-----|----|-----------------------------|--------------------------------|
| 10  | 0.1| 78.4002                     | 0.0235                         |
|     | 0.2| 61.4100                     | 0.0260                         |
|     | 0.3| 55.8004                     | 0.0210                         |
|     | 0.5| 50.1508                     | 0.0150                         |
| 20  | 0.1| 127.1000                    | 0.1235                         |
|     | 0.2| 103.4510                    | 0.0858                         |
|     | 0.3| 90.7502                     | 0.0855                         |
|     | 0.5| 81.8514                     | 0.0655                         |
| 30  | 0.1| 141.9500                    | 0.1780                         |
|     | 0.2| 126.4934                    | 0.1585                         |
|     | 0.3| 111.6512                    | 0.1325                         |
|     | 0.5| 95.0105                     | 0.1195                         |

From Table 1-3, for the different the value p, we observe that the average number of iterations and the average elapsed time decrease when the factor balance ρ increases. The average number of iterations and the average elapsed time decrease with the value ρ when the values of k and p are fixed. For example, when k=20 and ρ = 0.5 , the average number of iterations is 83.12 and the average elapsed time is 0.2143 when p=3, while the average number of iterations is 82.9808 and the average elapsed time is 0.2001 when p=5, and the average number of iterations is 81.8514 and the average elapsed time is 0.0655 when p=7. The average number of iterations and the average elapsed time increase with the value of k when ρ and p are fixed. For example, when ρ = 0.5 and p=3, the average number of iterations is 83.12 and the average elapsed time is 0.2143 when k=20, while the average number of iterations is 101.0484 and the average elapsed time is 0.3468 when k=30.

In this simulation study, we show that the optimal allocation that optimizes the objective function in equation (2) can be obtained based on Theorem 1. The simulation results show that the average number of iterations and the average elapsed time of the optimal allocation computation are reasonable small, which address the slow convergence problem.

We consider a numerical example for a set of particular values of p=3, k=10 and costs r is generated uniform distributed in between 0.001 and 1, (r_1, r_2, ..., r_{10}) = (0.84, 0.28, 0.79, 0.10, 0.34, 0.63, 0.08, 0.45, 0.67, 0.83), then we can obtain the optimal values (w_1, w_2, ..., w_{10}) = (0.036, 0.208, 1, 0.128, 0.471, 0.241, 0.139, 1, 1, 0.539) basing on the Theorem 1, 0.2 \sum_{i=1}^{10} r_i w_i = 0.5563 , and (r_1, r_2, ..., r_{10}) = (0.95, 0.5, 0.78, 0.23, 0.36, 0.43, 0.91, 0.24, 0.93, 0.89) the optimal values (w_1, w_2, ..., w_{10}) = (0.191, 0.728, 0.238, 0.051, 1, 0.289, 0.376, 1, 0.92, 1), 0.2 \sum_{i=1}^{10} r_i w_i = 0.7110, the cost is larger, but the precision is less, so we can show that the method balance between the cost and the efficiency in estimation.

5. Conclusions and future research direction
This paper deals with optimal experimental designs with cost consideration. An objective function that incorprorates the cost is proposed and the procedure to obtain the optimal allocation is presented. When the cost for running the experiments at different levels are the same, the results presented in this paper are the same as the results in reference [7]. One possible extension of the current work is to consider other kinds of regression model.

Acknowledgments
The work is supported by the National Natural Science funds (11471068).

References
[1] Fedorov V V, Gagnon R C and Leonov S L 2002 Appl. Stochastic Models Bus Ind. 18 207-218
[2] Fedorov V V and Leonov S L 2005 Response-Driven Designs in Drug Development, in *Applied Optimal Designs*, eds. M. Berger and W.-K. Wong (New York: Wiley)

[3] Wright S E, Sigal B M and Bailer A J 2010 *J. Agr Bio Envi Stat* 15 491-509

[4] Retout S, Comets E, Bazzoli C and Mentré F 2009 *Commun Stat–Theor M* 38 3351-3368

[5] Tekle F B, Tan F E S and Berger M P F 2008 *Stat. Med* 27 2586-2600

[6] Yu Y 2010 *Annu. Stat* 38 1593-1606

[7] Gao W, Chan P S, Ng H K T and Lu X 2014 *J. Comput Appl Math* 261 118-126

[8] Kuhn H W and Tucker A W 1951 Nonlinear programming, *Proceedings of 2nd Berkeley symposium* (California: Berkeley)