ESTIMATING THE RESTRICTED POPULATION PROPORTION USING Linear PROGRAMS

Ramadan Hamed Mohamed
Faculty of Economics & Political Sci.
Cairo University , Cairo , Egypt.

Abstract :

The problem of estimating the population proportion p_{ij} (proportion of individuals being in the i^{th} group of some factor and the j^{th} group of another one), when the marginal proportions of these factors are known a priori, was investigated and solved using nonlinear programming. In this paper , the linear goal programming and the linear programming are suggested to be used in this area. Using any of these suggested linear programs to solve the mentioned problem has a main advantage over the nonlinear programming

In nonlinear programming , either the sum of squares of deviations is minimized subject to the constraints of prior marginal proportions , or the likelihood function is maximized with the same restrictions . In both cases, the known difficulties and drawbacks of nonlinearity take place. One of these drawbacks is that we are not sure whether the solution we get , if we could , is an optimal , close to optimal, global , or local one and in most cases this solution depends on the initial value(s) . In the suggested linear approaches the sum of absolute deviations is minimized , subject to the same prior marginal proportions . Using these approaches is easier and many computer packages for solution are available . Moreover the parametric or sensitivity analysis of linear programming could be used to study the effects of the changes in the sample observations and/or the prior information on the estimates . An example for illustration is given in this paper . Also, this example is used to prove the ability of the suggested models to give correct solutions by considering a case of having a sample with the same population proportions .

Key Words : Linear Programming , Goal Programming , Estimation .

1. Introduction

Consider the case of having two (for simplicity) attributes y_1 and y_2 with k_1 and k_2 mutually exclusive classes , respectively . The probability of an individual , drawn randomly from the population , to be in the i^{th} class of y_1 $(i=1,2,...,k_1)$ and j^{th} class of y_2 $(j=1,2,...,k_2)$ is p_{ij} . Suppose that the p_{ij}'s are unknown and we need to estimate them . If we draw a random sample from the population with size n and we found that :

O_{ij} : the number of observations in the i^{th} class of y_1 and j^{th} class of y_2 $(i=1,2,...,k_1$, $j=1,2,...,k_2$), then the proportion O_{ij} / n is the maximum likelihood estimator of p_{ij} [8,9]. If in addition to the previous sample information we know a priori the marginal proportions of both y_1 and y_2 and that the two attributes are not independent, then we have the problem of estimating the population proportions p_{ij} under the restrictions of prior marginal proportions $p_{i.}$ for y_1 and $p_{.j}$ for y_2 ($i=1,2,\ldots,k_1$; $j=1,2,\ldots,k_2$).

The mentioned problem was investigated and solved using the nonlinear programming by minimizing the sum of weighted squares of deviations or by maximizing the likelihood function, under the prior marginal conditions [1,16]. In this paper the linear goal programming (LGP) and the linear programming (LP) are suggested as alternative approaches that have more advatanges over the nonlinear approaches . In the suggested approaches the exact optimal solutions are obtained and not close to optimal or approximated results . Also , the parametric and/or the sensitivity analysis of LP can be used to study the stability of the solution.

The two suggested approaches are based on the definition of deviational variable in LGP, therefore a background about the goal programming formulation is presented in section 2 of this paper , the suggested approaches for estimating p_{ij} are presented in section 3 , and , an example for illustration, is given in section 4 . The example is used also to prove the ability of the suggested approaches to estimate correct proportions if the drawn sample has the same population proportions .

2. Goal Programming (GP)

GP was firstly introduced by Charnes & Cooper [4] in the fifties as an approach to obtain estimates in the constrained regression . Since then, more and more modifications and applications were contributed to the approach [2,3,5,6,7,11] . GP is used mainly in the multicriteria decision making problems where there are several conflicting goals to be achieved .

To formulate a problem in a GP model , the following steps should be considered :

1. The first step is to assign an aspiration level for each objective, for example if we have the following multiobjective program:

$$\text{Opt. } f_i(X) \quad i=1,2,\ldots,M \tag{1}$$

s.t.

$$X \in F \tag{2}$$

where X is the decision variables vector , $f_i(X)$ is the i^{th} objective, M is the number of objectives , and F is the set of feasible solutions, then we should assign a target for each objective, let this target be b_i for the i^{th} objective. i.e.:

$f_i(X) < (>) b_i \qquad i=1,2,\ldots,M \qquad (3)$

2. Since we have conflicting goals, it is not expected to achieve the aspiration levels for all the goals. It is accepted to have deviations from the targets. These deviations are called negative deviations, n_i, if the realized objective is less than its target, and are called positive deviations, p_i, if the realized objective is greater than its target. These deviations are introduced into the GP model as follows:

$f_i(X) + n_i - p_i = b_i \qquad i=1,2,\ldots,M \qquad (4)$

Now our objective is to minimize the undesired deviation for each goal, i.e. if our original objective is to be maximized, we should minimize its positive deviation p_i and if another objective is to be minimized then its negative deviational variable should be minimized. In some cases we need $f_i(X) = b_i$ and in these cases n_i+p_i must be minimized.

3. Before establishing the GP model, the decision maker should assign a priority level for each goal, let q be the number of priority levels. The last step in GP formulation is to construct the achievement function (same as the objective function). There are many variants of achievement function in GP, the

weighted GP, the minmax GP, and the lexicographic GP are the most widely used forms . In this paper the later will be used which takes the following form :

$$\text{lexico min}A=\{h_1 (n_i,i\in p_i), h_2(n_i, i\in p_i)..., h_q(n_i, i\in p_i)\} \quad (5)$$

s.t.

$$f_i(X) + n_i - p_i = b_i \qquad i=1,2,...,M \quad (6)$$

$$X \in F \quad (7)$$

$$n_i , p_i > o \quad n_i\ p_i =0 \qquad i=1,2,....,M \quad (8)$$

where h(.) is a real valued function of the deviational variables , A is an ordered vector with dimension q (the number of priority levels) .

If $f_i(.)$ and $h_i(.)$ are linear functions and all the parameters of the GP (5)-(8) are constants and known , then the program is a LGP that can be solved by the sequential simplex method [6,7] .

GP was applied in many areas such as academic resources planning [11] , accounting [13] , quality control [14,17] , and many other applications. A good survey for these applications is in [15] .

3. Estimating p_{ij} using linear programs

3.1 statement of the problem :

If the two attributes y_1 and y_2 are independent, then :

$p_{ij} = p_{i.}\, p_{.j}$ (9)

but if this hypothesis is rejected, then we can not estimate p_{ij} with this equation, also using the maximum likelihood estimate from the sample may lead to values of p_{ij} that do not satisfy the conditions of prior marginal proportions. The problem now is how to determine the values of p_{ij} and these values should :

(1) satisfy the prior information , expressed by the following prior marginal constraints :

$$\sum_{j=1}^{k_2} p_{ij} = p_{i.} \qquad i=1,2,\ldots,k_1 \qquad (10)$$

$$\sum_{i=1}^{k_1} p_{ij} = p_{.j} \qquad j=1,2,\ldots,k_2 \qquad (11)$$

and

(2) take into account the posterior information from the sample , i.e. make the observed frequencies as close as possible from the expected frequencies . This objective can be expressed mathematically as :

$$\min \sum_{i=1}^{k_1}\sum_{j=1}^{k_2} \left| O_{ij} - np_{ij} \right| \qquad (12)$$

Up to this point , the problem of estimating p_{ij} is reduced to finding p_{ij} that satisfies (10) , (11) as system constraints and (12) as an objective .

3.2 The LGP Model :

The problem as stated in the subsection 3.1 was investigated and solved using nonlinear programming methods [1,16]. In most cases the solutions obtained using nonlinear programming are not global optimal and in some cases they are local or close to optimal . Using linear programs is always easier and gives optimal solutions . Therefore the following LGP is suggested to be used to solve the problem after refmulating it as follows:

Find p_{ij} ($i=1,2,\ldots,k_1$; $j=1,2,\ldots,k_2$) that lexicographically minimize A where:

$$A = \{ \sum_{i=1}^{k_1}(n_{i1} + n_{i2}) + \sum_{j=1}^{k_2}(s_{j1}+ s_{j2}); \ \sum_{i=1}^{k_1}\sum_{j=1}^{k_2} d_{ij} + t_{ij} \} \qquad (13)$$

s.t.

$$\sum_{j=1}^{k_2} p_{ij} + n_{i1} - n_{i2} = p_{i.} \quad i=1,2,\ldots,k_1 \qquad (14)$$

$$\sum_{i=1}^{k_1} p_{ij} + s_{j1} - s_{j2} = p_{.j} \quad j=1,2,\ldots.,k_2 \qquad (15)$$

$$np_{ij} + d_{ij} - t_{ij} = O_{ij} \quad i=1,2,\ldots,k1 , j=1,2,\ldots,k2 \qquad (16)$$

$$n_{i1}n_{i2} = 0 , \quad s_{j1}s_{j2}=0 , \quad d_{ij}\, t_{ij} =0 \quad i=1,2,\ldots,k1 , j=1,2,\ldots,k2 \qquad (17)$$

$$p_{ij}, \ n_{i1}, n_{i2}, \ s_{j1} , s_{j2} , d_{ij}, \text{ and } t_{ij} > 0 \quad i=1,2,\ldots,k1 , j=1,2,\ldots,k2 \qquad (18)$$

where :

1. n_{i1} , n_{i2} ($i=1,2,\ldots,k_1$) are the negative and positive deviational variables, respectively , of constraints (10) and minimizing n_{i1}+ n_{i2} is sufficient to satisfy these conditions .

2. s_{j1} , s_{j2} (j=1,2,...,k_2) are (similar as n_{i1} , n_{i2}) the negative and positive deviational variables, respectively , of constraints (11) and minimizing their sum is sufficient to satisfy these conditions .

The objective (12) is transformed to a goal (16) in the GP model with d_{ij} and t_{ij} the negative and positive deviational variables,respectively , of the goal constraint . Minimizing d_{ij} + t_{ij} is equivalent to satisfing (12) as proved in the following lemma .

Lemma:

The values of p_{ij} that make d_{ij}+t_{ij} minimum are the same values that satisfy (12) .

Proof:

since $d_{ij} = \max \{ 0 ; np_{ij} - O_{ij} \} = 0.5(np_{ij} - O_{ij} + |np_{ij}- O_{ij}|)$

and $t_{ij} = \max \{ 0 ; O_{ij} - np_{ij} \} = 0.5(O_{ij} - np_{ij} +|O_{ij} - np_{ij}|\}$

then $\sum_{i=1}^{k_1}\sum_{j=1}^{k_2}(d_{ij} + t_{ij}) =\sum_{i=1}^{k_1}\sum_{j=1}^{k_2}|O_{ij} - np_{ij}|$

So, satisfing (12) is equivalent to minimizing $\sum_{i=1}^{k_1}\sum_{j=1}^{k_2}(d_{ij}+ t_{ij})$

Now, the system constraints (10) and (11) are satisfied by (14) and (15) in the program and they are represented in the achievement function A in the first priority level , and the objective (12) is represented by the goals (16) in the LGP , these goals are ranked in the second priority level of the achievement function A.

The model (13)- (18) is a linear goal program that can be solved by any of the multiphase simplex or the sequential goal programming methods [5,6].

3.2 The LP Model :

The LGP (13)-(18) can be converted to an equivalent LP using the following steps:

1. canceling the deviational variables of system constraints .
2. The achievement function will contain only the deviational variables of the second priority goal and hence it becomes a single objective function .

The problem under investigation can be solved as a LP which takes the following form :

$$\text{Min } Z = \sum_{i=1}^{k_1}\sum_{j=1}^{k_2} (d_{ij} + t_{ij}) \qquad (19)$$

s.t.

$$\sum_{j=1}^{k_2} p_{ij} = p_{i.} \qquad i=1,2,\ldots,k_1 \qquad (20)$$

$$\sum_{i=1}^{k_1} p_{ij} = p_{.j} \qquad j=1,2,\ldots.,k_2 \qquad (21)$$

$$np_{ij} + d_{ij} - t_{ij} = O_{ij} \qquad i=1,2,\ldots,k1 \text{ , } j=1,2,\ldots,k2 \qquad (22)$$

$$p_{ij},\ n_{i1},\ n_{i2},\ s_{j1},\ s_{j2},\ d_{ij}, \text{ and } t_{ij} > 0 \qquad (23)$$

It is easy to prove the equivalence between the LGP and the LP if the later is feasible by using the relationship between LGP and LP mentioned in many references as [6], but if the LP has infeasible solution, then one should use the LGP, and that is why the two approaches are included in this paper.

4. A Numerical Example

In December 1993, the first national census in the Sultanate of Oman was carried out. The relative distribution of the population by nationality (I) and region (j), p_{ij}, is given in table (1). Suppose that we do not know the joint relative distribution and we know only the distribution of the population by nationality and the distribution by region. To estimate the joint relative distribution of the population by the two attributes, let us draw a random sample of size 10000 individuals, and suppose that the results of this sample are as shown in table (2).

Table (1)

The relative distribution of the population by nationality and region , p_{ij}.

Region	Muscat	Dhofar	Dakhelya	Sharqiya	Batinah	Other	Total
Omani	.163	.060	.096	.106	.227	.082	.734
Nonomani	.145	.027	.013	.017	.040	.024	.266
Total	.308	.087	.109	.123	.267	.106	1

Table (2)

The observed frequencies of the sample , O_{ij}.

Region	Muscat	Dhofar	Dakhelya	Sharqiya	Batinah	Other	Total
Omani	1800	560	1000	1110	2500	1000	7970
Nonomani	810	350	100	210	410	150	2030
Total	2610	910	1100	1320	2910	1150	10000

Now to estimate p_{ij} using only $p_{i.}$, $p_{.j}$, and O_{ij} we should first test the independence of the two attributes region and nationality and by using the Chi-Squared test the hypothesis of independence is rejected (the calculated χ^2 =579.366 compared withthe tabulated χ^2 =15.086 at 0.01 significance level) and so we can not find p_{ij} by multiplying $p_{i.}$ by $p_{.j}$.

We can use either the LGP or LP to obtain the estimates of P_{ij} (i=1,2 j=1,2,3,4,5,6) , as follows :

(1) By using the LGP (13) - (18) , the marginal proportions from table (1) , and the observed frequencies from the sample as shown in table (2) , we can get the following LGP model :

$$\text{Lexico min } A = \{\sum_{i=1}^{2} (n_{i1}+n_{i2})+\sum_{j=1}^{6}(s_{j1}+s_{j2}), \quad \sum_{i=1}^{2}\sum_{j=1}^{6}(d_{ij} + t_{ij})\} \quad (24)$$

s.t.

The conditions of prior marginal : (25)

$$\sum_{j=1}^{6} p_{1j} + n_{11} - n_{12}= .734 \quad , \text{ and } \sum_{j=1}^{6} p_{2j} + n_{21} - n_{22}= .266$$

$$\sum_{i=1}^{2} p_{i1} + s_{11} - s_{12} = .308 \quad , \quad \sum_{i=1}^{2} p_{i2} + s_{21} - s_{22}= .087$$

$$\sum_{i=1}^{2} p_{i3} +s_{31} - s_{32}= .109 \quad , \quad \sum_{i=1}^{2} p_{i4} + s_{41} - s_{42}= .123$$

$$\sum_{i=1}^{2} p_{i5} + s_{51} - s_{52}= .267 \quad , \quad \sum_{i=1}^{2} p_{i6} + s_{61} - s_{62}= .106$$

,The conditions of posterior information from the sample : (26)

$$10000\, p_{11} + d_{11} - t_{11} = 1800 \quad , \quad 10000p_{21}+ d_{21} - t_{21} = 810 ,$$

$$10000\, p_{12} + d_{12} - t_{12} = 560 \quad , \quad 10000p_{22} + d_{22} - t_{22} = 350 ,$$

$$10000\, p_{13} + d_{13} - t_{13} = 1000 \quad , \quad 10000p_{23} + d_{23} - t_{23} = 100 ,$$

$$10000\, p_{14}+ d_{14} - t_{14}= 1110 \quad , \quad 10000p_{24}+ d_{24} - t_{24} = 210 ,$$

$$10000\, p_{15}+ d_{15} - t_{15} = 2500 \quad , \quad 10000p_{25}+ d_{25} - t_{25} = 410 ,$$

$$10000\, p_{16} + d_{16}- t_{16} = 1000 \quad , \text{ and } 10000p_{26}+ d_{26} - t_{26} = 150 ,$$

and the nonnegativity conditions : (27)

$$p_{ij} , n_{i1} , n_{i2} , s_{j1} , s_{j2} , d_{ij} , t_{ij} > 0$$

$$n_{i1}\, n_{i2} = 0 , \; s_{j1}s_{j2} = 0 \; , \; d_{ij}\, t_{ij} = 0 \quad I=1,2 \; , \; j=1,2,3,4,5,6$$

Also, by using the LP (19) - (23) together with the same input data about marginal and observed frequencies , we can get the following LP:

$$\text{Min } Z = \sum_{i=1}^{2}\sum_{j=1}^{6} d_{ij} + t_{ij} \qquad (28)$$

s.t.

The conditions of prior marginal : (29)

$\sum_{i=1}^{2} p_{i1} = .308$, $\sum_{i=1}^{2} p_{i2} = .087$, $\sum_{i=1}^{2} p_{i3} = .109$, $\sum_{i=1}^{2} p_{i4} = .123$

$\sum_{i=1}^{2} p_{i5} = .267$, $\sum_{i=1}^{2} p_{i6} = .106$, $\sum_{j=1}^{6} p_{1j} = .734$, and $\sum_{j=1}^{6} p_{2j}.266$

,The conditions of posterior information from the sample : (30)

$10000\, p_{11} + d_{11} - t_{11} = 1800$, $10000p_{21} + d_{21} - t_{21} = 810$,

$10000\, p_{12} + d_{12} - t_{12} = 560$, $10000p_{22} + d_{22} - t_{22} = 350$,

$10000\, p_{13} + d_{13} - t_{13} = 1000$, $10000p_{23} + d_{23} - t_{23} = 100$,

$10000\, p_{14} + d_{14} - t_{14} = 1110$, $10000p_{24} + d_{24} - t_{24} = 210$,

$10000\, p_{15} + d_{15} - t_{15} = 2500$, $10000p_{25} + d_{25} - t_{25} = 410$,

, and the nonnegativity conditions : (31)

p_{ij}, , d_{ij} , $t_{ij} > 0$ I=1,2 , j=1,2,3,4,5,6

By using the Micro Manager Package [12] , the optimal solutions of both the LGP (24) - (27) and the LP (28) - (31) are given in table (3) . In table (4) the lower and upper bounds of the observed frequencies (O_{ij}) that can be used for the sensitivity analysis of the solution , are stated .

Table (3)

Estimates of p_{ij} from the LGP and LP

Region	Muscat	Dhofar	Dakhelya	Sharqiya	Batinah	Other	Total
Omani	0.164	.052	.099	.102	.226	.091	.734
Nonomani	.144	.035	.01	.021	.041	.015	.266
Total	.308	.087	.109	.123	.267	.106	1

Table (4)

Bounds of Parameters O_{ij}

O_{ij}	Current value	Upper limit	Lower limit
O_{11}	1800	No	1640
O_{21}	810	1440	No
O_{12}	560	NO	520
O_{22}	350	510	310
O_{13}	1000	NO	990
O_{23}	100	260	90
O_{14}	1110	NO	1020
O_{24}	210	370	120
O_{15}	2500	NO	2260
O_{25}	410	570	170
O_{16}	1000	No	910
O_{26}	150	310	60

Now, suppose we have a sample with the same proportions of the population as given in table (1), could the suggested models give the same proportions as solutions ? . To answer, let us reformulate the models (24) - (27) and (28) - (31) by using a new suggested observed frequencies $O_{ij} = 10000\ p_{ij}$ (I=1,2 , j=1,2,3,4,5,6) where p_{ij} are as defined in table (1). These values will replace the right hand side values of the posterior conditions (26) and (30) . Hence , the optimal solutions as given in Table (5) show that the estimates of the proportions are as the same values as the proportions of the population . This proves the ability of the suggested models to find the correct estimates of the proportions .

Table (5)

Estimates of p_{ij} from the LGP and LP

Region	*Muscat*	*Dhofar*	*Dakhelya*	*Sharqiya*	*Batinah*	*Other*	*Total*
Omani	*.163*	*.060*	*.096*	*.106*	*.227*	*.082*	*.734*
Nonomani	*.145*	*.027*	*.013*	*.017*	*.040*	*.024*	*.266*
Total	*.308*	*.087*	*.109*	*.123*	*.267*	*.106*	*1*

5. Conclusion

The problem of estimating the population proportions of two attributes when the marginal proportions of these attributes are known a priori , is investigated in this paper and new approaches for estimation are developed . These approaches depend on formulating the problem as a LGP or a LP program . Using any of the two models will give the global optimal solution and enables us to make a sensitivity analysis and see to what extent the changes in the sample observations will affect the solutions . The statistical properties of the estimators given by these approaches are not studied in this paper . The author intends to investigate these properties in the very near future .

REFERENCES

[1] Arthanari, T. S. and Yadolah Dodge. (1981). *Mathematical programming in Statistics.* John Wily & Sons, New York.

[2] Charnes, A. and Cooper, W. W. (1961). *Management models and industrial applications of linear programming.* John Wily & Sons New York.

[3] Charnes, A. and Cooper, W. W. (1977). Goal programming and multiple objective optimization. Part I. European *Journal of Operational Research.* 1, 39-54.

[4] Charnes, A., and Cooper, W. W., and Ferguson, R. (1955). Optimal estimation of executive compensation by linear

programming, *Management Science*, 1, 138-151.

[5] Ignizio, J. P. (1976). *Goal programming and its extension.* Lexington Lexington Books, Massachusetts.

[6] Ignizio, J. P. (1982). *Linear programming in single and multiple-objective systems.* Prentice-Hall, Englewood Cliffs.

[7] Ignizio, J. P. (1983). Generalized goal programming. An overview. *Computers & Operations Research*, 10, 277-289.

[8] Kendall, M. G. and Stuart, A. (1991). *The advanced theory of statistics.* Charles Griffin, London.

[9] Kullback, S. (1974). Loglinear models in contingency table analysis. *An. Stat.* 28, 115-125.

[10] Lee, S. M. (1972). *Goal programming for decision analysis.* Auerbach publishers, Philadelphia.

[11] Lee, S. M. and Clayton, E. R. (1972). A goal programming model for academic resource allocation, *Management Science*, 18, 395-408 .

[12] Lee, S. M. and Shim, J. P. (1986) *Micro management science.* Wm. C. Brown Publishers, Dubuque, Iowa.

[13] Lee, S. M. and Eom, H. B. (1989) A multi-criteria approach to formulating international project-financing strategies. *Journal of Operational Research Society*, 40, 519-528.

[14] Ravindran, A., Shin W. S. Arthur, J. L. and Moskowitz, H. (1986).Nonlinear integer goal programming models for acceptance sampling. *Computer and Operations Research*, 13, 611-622.

[15] Romero, C. (1991). *Handbook of critical issues in goal programming*. Pergamon Press, Oxford.

[16] Rustagi, J. S. Ed. (1979). *Optimizing methods in Statistics.* II. Academic Press, New York.

[17] Sengupta, S. (1981). Goal programming approach to a type of quality control problem. *Journal of the Operational Research Society*, 32, 207-211.