A NOTE ON PREDICTION IN THE CASE OF FINITE POPULATIONS

By

K.C. Gautam and Padam Singh

I.A.S.R.I., New Delhi-110012

(Received : June, 1976)

1. Introduction

Regression relations are of considerable importance for predicting the value of one variable on the basis of given value of an auxiliary variable. In the literature on regression analysis it is assumed that the population under consideration is infinite which is not the case always. In practice, the populations under study are usually finite. In the present paper, we study the general problem of regression analysis when the population is finite. The expressions for the variance of the regression predictor have been worked out for various situations depending on how the value of the auxiliary variate is chosen.

2. Main Results

Consider a finite population consisting of N units and let y and x be two variables taking values y_j and x_j for the j^{th} unit of the population, $j=1,\ldots\ldots, N$. A relationship of the type $y=\alpha+\beta x$ is desired to be determined for the population. For this, let a sample of size 'n' be drawn and let $(x_1, y_1), (x_2, y_2),\ldots\ldots, (x_n, y_n)$ be the pairs of values of x and y respectively on 'n' units of the population. The procedure of fitting the regression equation consists of estimating α and β such that the sum of squares of the deviations from the line of regression is minimum. Following least square technique, the estimates of α and β on the basis of sample are given by

$$\hat{\beta}=b=\frac{\sum_{j=1}^{n}(x_j-\bar{x}_n)(y_j-\bar{y}_n)}{\sum_{j=1}^{n}(x_j-\bar{x}_n)^2} \qquad \ldots(2.1)$$

and

$$\hat{\alpha}=a=\bar{y}_n-\hat{\beta}\bar{x}_n \qquad \ldots(2.2)$$

where $\bar{x}_n$ and $\bar{y}_n$ are the sample means of characters x and y respectively.

Having obtained the estimates of α and β by least square technique, the next aspect of regression analysis is to predict the value of y corresponding to some value of x. The value of x may be any one of the following :

(*a*) Some given x-value which is decided subjectively.

(*b*) The x-value of a unit selected from the sample.

(*c*) The x-value of a unit selected from the units not in the sample.

(*d*) The x-value of a unit chosen from the whole population.

There is considerable literature relating to the situation (*a*) such as response to given level of fertilizer, milk yield for a given intake, output for given input etc. There are other types of situations such as prediction of yield corresponding to plant population or plant height, prediction of supply corresponding to price level etc. where plant population, plant height and the price level themselves are random variables. We shall examine the unbiasedness of the predictor and the error associated with it in the last three situations mentioned above. However, for clarity and completeness the case corresponding to situation (*a*) has also been included here.

2.1. **Case** (*a*) : This pertains to the situation where the unit for which the value of y has to be predicted is selected subjectively. It is also assumed that x_j's are fixed constants and not random variables. The assumption in the prediction model

$$y_j=\alpha+\beta x_j+\epsilon_j, (j=1,\ldots\ldots, N) \qquad \ldots(2.3)$$

is that ϵ_j's are independently normally distributed with zero mean and constant variance σ_e^2.

In this case $\hat{\beta}$, given by (2.1) is the least square estimate of β and the variance of the predictor $\hat{Y}_j$ is

$$V(\hat{Y}_j)=\frac{\sigma_e^2}{n}+\frac{(x_j-\bar{x}_n)^2\,\sigma_e^2}{\sum_{j=1}^{n}(x_j-\bar{x}_n)^2} \qquad \ldots(2.4)$$

This is minimum when $x_j=\bar{x}_n$ and increases as x_j moves away from $\bar{x}_n$ in either direction.

2.2. **Case** (*b*): Consider the situation when x is the value of a unit selected from the units in the sample by simple random sampling.

We have the regression equation

$$y_j=\bar{y}_n+b(x_j-\bar{x}_n) \qquad ...(2.5)$$

The predictor $\hat{Y}_j$ of y corresponding to x_j, $(j=1\ldots\ldots, n)$ is given by

$$\hat{Y}_j=\bar{y}_n+b(x_j-\bar{x}_n) \qquad ...(2.6)$$

The difference between the true value and the predictor corresponding to j^{th} unit is

$$y_j-\hat{Y}_j=y_j-\bar{y}_n-b(x_j-\bar{x}_n) \qquad ...(2.7)$$

The expectation of the above expression has to be taken twice, first for the given sample and then for all possible samples. Thus in the usual notation, we have

$$\begin{aligned}E(y_j-\hat{Y}_j)&=E_1E_2[y_j-\bar{y}_n-b(x_j-\bar{x}_n)]\\&=E_1[\bar{y}_n-\bar{y}_n-b(\bar{x}_n-\bar{x}_n)]=0\end{aligned} \qquad ...(2.8)$$

since $E_2(x_j)=\bar{x}_n$, $E_2(y_j)=\bar{y}_n$ because of selection and hence the predictor $\hat{Y}_j$ is unbiased.

Now, the variance of prediction with usual notations is given by

$$\begin{aligned}V(\hat{Y}_j)&=E_1V_2(\hat{Y}_j)+V_1E_2(\hat{Y}_j)\\&=E_1\frac{n-1}{n}(1-r^2)\,s_y^2\\&=\frac{n-1}{n}\sum_{s\in S}p_s\left(s_y^2-\frac{s_{xy}^2}{s_x^2}\right)\end{aligned} \qquad ...(2.9)$$

where p_s is the probability of selecting the sample and S denotes the totality of all possible samples of size n.

For the simple situation, when the original sample has been drawn employing simple random sampling without replacement, the variance of the predictor is given by

$$V(\hat{Y}_j)=\frac{n-1}{n}\left[S_y^2-E\frac{s_{xy}^2}{s_x^2}\right] \qquad ...(2.10)$$

Let

$$s_{xy}=S_{xy}+\epsilon_{11}$$

$$s_x^2=S_x^2+\epsilon_{01}$$

We have

$$E\left(\frac{s_{xy}^2}{s_x^2}\right) = E\frac{(S_{xy}+\epsilon_{11})^2}{S_x^2+\epsilon_{01}}$$

$$= \frac{S_{xy}^2}{S_x^2} E\left(1+\frac{\epsilon_{11}}{S_{xy}}\right)^2\left(1+\frac{\epsilon_{01}}{S_x^2}\right)^{-1}$$

Next, assuming $\left|\frac{\epsilon_{01}}{S_x^2}\right|<1$, which is sufficiently justified, we can expand $\left(1+\frac{\epsilon_{01}}{S_x^2}\right)^{-1}$ to a suitable number of terms and then taking expectations we get

$$E\frac{s_{xy}^2}{s_x^2} = \frac{\mu_{11}^2}{\mu_{02}}\left\{1+\frac{V(m_{11})}{\mu_{11}^2}-\frac{2\,Cov\,(m_{11}, m_{02})}{\mu_{11}\mu_{02}}+\frac{V(m_{02})}{\mu_{02}^2}\right\} \quad ...(2\cdot11)$$

Further, we know that (Kendall and Stuart (1969), page 235)

$$V(m_{11})=\frac{1}{n}\left(\mu_{22}-\mu_{11}^2\right) \quad ...(2\cdot12)$$

$$Cov\,(m_{11}, m_{02})=\frac{1}{n}(\mu_{13}-\mu_{11}\,\mu_{02}) \quad ...(2\cdot13)$$

$$V(m_{02})=\frac{1}{n}\left(\mu_{04}-\mu_{02}^2\right) \quad ...(2\cdot14)$$

On substituting the values of various terms, the variance of the predictor is given by

$$V_p=\frac{n-1}{n}\left[\mu_{20}-\frac{\mu_{11}^2}{\mu_{02}}-\frac{1}{n\mu_{02}}\left\{\mu_{22}-\frac{2\mu_{11}\mu_{13}}{\mu_{02}}+\frac{\mu_{11}^2\,\mu_{04}}{\mu_{02}^2}\right\}\right] \quad ...(2\cdot15)$$

which, to a first degree of approximation, takes the form

$$V_p\cong\mu_{20}-\frac{\mu_{11}^2}{\mu_{02}}$$

2.3 Case C: Now, let us consider the case when x is the value of a unit selected from the remaining $(N-n)$ units of the population. We have a regression equation

$$y_j=\bar{y}_n+b(x_j-\bar{x}_n)$$

The predicted value of y corresponding to x_j, $j=n+1,\ldots, N$ is given by

$$\hat{Y}_j=\bar{y}_n+b(x_j-\bar{x}_n)$$

Now the difference between the true value and the predictor $\hat{Y}_j$ is given by

$$y_j-\hat{Y}_j=y_j-\bar{y}_n-b(x_j-\bar{x}_n)$$

Also, since jth unit has been selected from the remaining $(N-n)$ units we have

$$E_2(x_j)=\bar{X}_{N-n}=\frac{N\bar{X}_{N-n}\,\bar{x}_n}{N-n}=\frac{N}{N-n}\bar{X}_N-\frac{n}{N-n}\bar{x}_n$$

and

$$E_2(y_j)=\bar{Y}_{N-n}=\frac{N}{N-n}\bar{Y}_N-\frac{n}{N-n}\bar{y}_n$$

Thus

$$E(y_j-\hat{Y}_j)=E_1\left[\frac{N}{N-n}\bar{Y}_N-\frac{n}{N-n}\bar{y}_n-\bar{y}_n-b\left(\frac{N}{N-n}\bar{X}_N-\frac{n}{N-n}\bar{x}_n-\bar{x}_n\right)\right]$$

$$=-\frac{N}{N-n}E_1\left[\bar{y}_n-\bar{Y}_N-b(\bar{x}_n-\bar{X}_N)\right]=0$$

and hence the predictor $\hat{Y}_j$ is unbiased.

Following the algebra of section (2.2), the variance of the predictor is given by

$$V_p=\frac{n-1}{n}\left[\mu_{20}-\frac{\mu_{11}^2}{\mu_{02}}-\frac{1}{n\mu_{02}}\left(\mu_{22}-\frac{2\mu_{11}\mu_{13}}{\mu_{02}}+\frac{\mu_{11}^2\,\mu_{04}}{\mu_{02}^2}\right)\right]+\left(\frac{1}{n}-\frac{1}{N}\right)\mu_{20}\,(1-\rho^2)\left(\frac{N}{N-n}\right)^2 \qquad \ldots(2{\cdot}16)$$

which, as a first degree of approximation, simplifies to

$$V_p\cong\delta\frac{N}{N-n}\cdot\frac{1}{n}\,(1-\rho^2)\,\mu_{20}+\left(\mu_{20}-\frac{\mu_{11}^2}{\mu_{02}}\right)$$

2.4. **Case *d*:** Finally consider the situation when x is the value of a unit selected from the whole population, the predictor $\hat{Y}_j$ corresponding to x_j, $(j=1,\ldots, N)$, is given by

$$\hat{Y}_j=\bar{y}_n+b(x_j-\bar{x}_n)$$

Now, since the unit under section is drawn from the whole population we have

$$E_2(x_j)=\bar{X}_N$$

and

$$E_2(y_j)=\bar{Y}_N$$

Thus, the difference between the true value and the predictor $\hat{Y}_j$ is given by

$$y_j-\hat{Y}_j=y_j-\bar{y}_n-b(x_j-\bar{x}_n)$$

and

$$E(y_j-\hat{Y}_j)=E_1[\bar{Y}_N-\bar{y}_n-b(\bar{X}_N-\bar{x}_n)]=0$$

hence the predictor is unbiased. Also, in this case, the variance of the predictor $\hat{Y}_j$ is given by

$$V_p=\left(\frac{1}{n}-\frac{1}{N}\right)(1-\rho^2)\,\mu_{20}+\frac{n-1}{n}\left[\mu_{20}-\frac{\mu_{11}^2}{\mu_{02}}-\frac{1}{n\mu_{02}}\left(\mu_{22}-\frac{2\mu_{11}\mu_{13}}{\mu_{02}}+\frac{\mu_{11}^2\,\mu_{04}}{\mu_{02}^2}\right)\right] \quad \ldots(2\cdot17)$$

which, to a first degree of approximation, takes the form

$$V_p\cong\left(\frac{1}{n}-\frac{1}{N}\right)(1-\rho^2)\,\mu_{20}+\left(\mu_{20}-\frac{\mu_{11}^2}{\mu_{02}}\right)$$

From (2.15), (2.16) and (2.17), it can be seen that the variance of predictor $\hat{Y}_j$ for cases (b), (c) and (d), obey the following inequalities,

$$V_p \text{ (Case } b)\leq V_p \text{ (Case } d) \leq V_p \text{ (Case } c).$$

SUMMARY

In this paper the problem of prediction from finite populations has been considered from various situations. It has been observed that there is a component in variance of prediction which depends upon the sampling procedure used for the selection of the sample.

REFERENCES

[1] Adcock, R J. (1878) : 'A problem on least square'. *The Analyst*, 5.

[2] Allen, R.G.D. (1939) : 'Assumptions of linear regression'. *Econometrica*, 6.

[3] Johnston, J. (1963) : *Econometric Methods*. McGraw Hill Book Co.

[4] Murthy, M.N. (1967) : 'Sampling Theory and Methods'. Statistical Publishing Society, Calcutta-35, India.

[5] Kendall, M.G. and Stuart, A. (1969) : *The Advanced Theory of Statistics*, Vol. 1'. Charles Griffin and Company Limited, 42, Durby Lane, London, W.C. 2.

[6] Sukhatme, P.V. and Sukhatme, B.V. (1972) : *Theory of Sampling with Applications*. Asia Publishing House, New Delhi-12.

[7] Williums, E.J. (1959) : *Regression Analysis*. New York, John Wiley and Sons.