PARABOLIC REGRESSION ESTIMATION

D.C. Doss and J. Hoomani, University of Alabama in Huntsville
G.R. Somayajulu*, Osmania University

## 1. INTRODUCTION

In this paper we estimate the population mean $\bar{Y}$ from a simple random sample $(y_1, y_2, \ldots, y_n)$ of size n. Suppose that the variable $y_i$ is related to an auxiliary variable, say $x_i$, where $\bar{X}$ is known. It is known that the linear regression estimator $\bar{y}_\ell = \bar{y} + b(\bar{X} - \bar{x})$ is unbiased provided that the target population is essentially infinite and the paired observations $(x_i, y_i)$ follow the usual linear model. Situations arise, however, in which while the assumption of an essentially infinite population is plausible, the model linking the variables $x_i$ and $y_j$ is of a quadratic form. Hence we propose the parabolic model

$$y_i = \beta_1 + \beta_2 x_i + \beta_3 x_i^2 + e_i, \quad i = 1,2,\ldots,n \tag{1.1}$$

Our assumptions are as follows: (a) The underlying population is very large, i.e., essentially infinite; (b) $\bar{X}$ is known; (c) $\beta_1$, $\beta_2$, $\beta_3$ are unknown parameters; (d) $e_1, e_2, \ldots, e_n$ are independent random variables and conditional expectations $E(e_i|x_i) = E(e_i) = 0$, $E(e_i^2|x_i) = \sigma_e^2$ for $i = 1,2,\ldots,n$; and finally (e)$x_1, \ldots, x_n$ are observable without error. Taking expectations of the model in (1.1) over the entire population, we note that

$$\bar{Y} = \beta_1 + \beta_2 \bar{X} + \beta_3 E(x^2). \tag{1.2}$$

Using the method of least square estimation, one can find unbiased estimates for $\beta_i$; however since population data on x are not generally available, $E(x^2)$ has to be estimated. Therefore, the parabolic estimator $\bar{y}_p$ of $\bar{Y}$ is proposed as

$$\bar{y}_p = \hat{\beta}_1 + \hat{\beta}_2 \bar{X} + \hat{\beta}_3 (\bar{X}^2 + s_x^2), \tag{1.3}$$

where $s_x^2 = \frac{1}{n-1} \Sigma_{i=1}^n (x_i - \bar{x})^2$, $(\bar{X}^2 + s_x^2)$ is an unbiased estimator of $E(x^2)$, and $\hat{\beta}_i$ is the least square estimator of $\beta_i$ for i = 1,2, 3. The purpose of this paper is to show that $\bar{y}_p$ is an unbiased estimator of the population mean $\bar{Y}$; to compute the exact variance $V(\bar{y}_p)$ of the parabolic estimator; and to determine an unbiased estimator $v(\bar{y}_p)$ for $V(\bar{y}_p)$.

## 2. UNBIASED ESTIMATOR OF $\beta_i$

We will use the subscript 1 to denote conditional expectation and variance given $x_i$, e.g., $E_1(Z) = E(Z|x_i)$.

Let

$A = \text{col}(y_1, \ldots, y_n)$,
$B = \text{col}(\beta_1, \beta_2, \beta_3)$,
$F = \text{col}(e_1, \ldots, e_n)$, and

$$P = \begin{bmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \cdots & \cdots & \cdots \\ 1 & x_n & x_n^2 \end{bmatrix}.$$

Then the parabolic model (1.1) becomes A = PB + F, or

$$F = A - PB \tag{2.1}$$

Since $\Sigma_{i=1}^n e_i^2 = F'F = (A - PB)'(A - PB)$, we are then to determine a vector $\hat{B} = \text{col}(\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3)$ that minimizes $\Sigma_{i=1}^n e_i^2$. Using the least square methods in regression analysis [2], we find that a solution of $\hat{B}$ is given by $\hat{B} = (P'P)^{-1}P'A$ and the conditional expectation of $\hat{B}$, given $x_i$, is $E_1(\hat{B}) = B$. Since the conditional expectation $E_1(A)$ of A is $E_1(A) = PB$, hence the unconditional expectation of B is $E(\hat{B}) = E[E_1(\hat{B})] = B$. Therefore $\hat{B} = (P'P)^{-1} P'A$ is an unbiased estimator of B.

## 3. EXPECTATION AND VARIANCE OF $\bar{y}_p$

Consider the expression (1.3) for $\bar{y}_p$. Taking conditional expectations of both sides, given $x_i$, we get

$$E_1(\bar{y}_p) = \beta_1 + \beta_2 \bar{X} + \beta_3(\bar{X}^2 + s_x^2).$$

Therefore

$$E(\bar{y}_p) = E[E_1(\bar{y}_p)] = \beta_1 + \beta_2 \bar{X} + \beta_3(\bar{X}^2 + S_x^2) = \bar{Y},$$

which shows that $\bar{y}_p$ is unbiased for $\bar{Y}$.

In order to compute the exact variance $V(\bar{y}_p)$, we first note that the conditional dispersion matrix $D_1$ of $\hat{B}$, given $x_i$, is obtained as follows.

$$D_1(P'P\hat{B}) = D_1(P'A)$$
$$= P'D_1(A)P = P'\sigma_e^2 IP.$$

Then

$$(P'P)D_1(\hat{B}) = \sigma_e^2 I$$

and

$$D_1(\hat{B}) = \sigma_e^2(P'P)^{-1}.$$

If we let

$$L = \mathrm{col}(1, \bar{X}, \bar{X}^2 + s_x^2),$$

we get

$$\bar{y}_p = L'\hat{B} = \hat{B}'L$$

and

$$V_1(L'\hat{B}) = L'D_1(\hat{B})L$$
$$= L'\sigma_e^2 (P'P)^{-1}L.$$

Hence

$$V_1(\bar{y}_p) = \sigma_e^2 L' (P'P)^{-1}L.$$

Finally, the unconditional variance becomes

$$V(\bar{y}_p) = E[V_1(\bar{y}_p)] + V[E_1(\bar{y}_p)]$$
$$= \sigma_e^2 E_x [L'(P'P)^{-1}L] + V(\beta_1 + \beta_2 \bar{X} + \beta_3 \bar{X}^2 + \beta_3 s_x^2).$$ Hence

$$V(\bar{y}_p) = \sigma_e^2 E_x [L'(P'P)^{-1}L] + \beta_3^2 V(s_x^2). \quad (3.1)$$

## 4. ESTIMATION OF $V(y_p)$

Let

$$R_o^2 = (A - P\hat{B})' (A - P\hat{B}),$$

which reduces to

$$R_o^2 = A'[I - P(P'P)^{-1}P']A,$$

and it can be shown [2] that

$$E_1(R_o^2) = (n-3)\sigma_e^2.$$

Moreover,

$$E(R_o^2) = (n-3)\sigma_e^2.$$

Therefore, an unbiased estimator $\hat{\sigma}_e^2$ of $\sigma_e^2$ is $R_o^2/(n-3)$. Next, we let

$$\hat{\beta_3^2} \equiv \hat{\beta}_3^2 - v(\hat{\beta}_3),$$

where $v(\hat{\beta}_3)$ represents an unbiased estimator of $V_1(\hat{\beta}_3)$. Since $D_1(\hat{B}) = \sigma_e^2(P'P)^{-1}$ we have

$$\hat{D}_1(\hat{B}) = \hat{\sigma}_e^2 (P'P)^{-1} \equiv \hat{\sigma}_e^2 [c_{ij}].$$

Then the element $v_{33}$ in the (3,3) position of $\hat{D}_1(\hat{B})$ is

$$v_{33} = \hat{\sigma}_e^2 c_{33}$$
$$= c_{33} R_o^2/(n-3). \quad (4.1)$$

We now note that

$$\hat{\beta_3^2} = \hat{\beta}_3^2 - c_{33} R_o^2/(n-3)$$

and

$$E_1(\hat{\beta_3^2}) = E_1(\hat{\beta}_3^2) - \sigma_e^2 c_{33}$$
$$= E_1(\hat{\beta}_3^2) - V_1(\hat{\beta}_3)$$
$$= [E_1(\hat{\beta}_3)]^2 = \beta_3^2,$$

which establishes that

$$\hat{\beta_3^2} \equiv \hat{\beta}_3^2 - c_{33} R_o^2/(n-3) \quad (4.2)$$

is an unbiased estimator of $\beta_3^2$.

To find an unbiased estimator of $V(s_x^2)$, we proceed as follows. Since

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x})^2$$
$$= \frac{n}{n-1} m_2,$$

where $m_2$ is the second moment about the sample mean, we can write the variance of $s_x^2$ in terms of the cumulants of x, [3].

$$V(s_x^2) = \left(\frac{n}{n-1}\right)^2 V(m_2)$$
$$= \frac{\kappa_4}{n} + \frac{2\kappa_2^2}{n-1}, \quad (4.3)$$

where $\kappa_r$ is the rth cumulant of x. Using the theory of symmetric functions [3], we recall that the k-statistic of order $k_r$, which is a function of the observations, is an unbiased estimator of $\kappa_r$ for all integers r. Moreover, the $\ell$-statistic (or polykay) $\ell_{22}$ has the property $E(\ell_{22}) = \kappa_2 \cdot \kappa_2 = \kappa_2^2$, where $\ell_{22}$ is also a function of observations only. Hence. if we let

$$\hat{V}(s_x^2) \equiv k_4/n + 2\ell_{22}/(n-1), \quad (4.4)$$

we get, from the tables of k - statistics,

$$\hat{V}(s_x^2) = \frac{1}{n+1} [\frac{n-1}{n} k_4 + 2k_2^2], \qquad (4.5)$$

where $k_2 = f(t_1, t_2)$, $k_4 = g(t_1, t_2, t_3, t_4)$

and $t_j = \sum_{i=1}^{n} x_i^j$.

Finally, it follows that

$$\hat{V}(\bar{y}_p) = \frac{1}{n-3} R_o^2 [L' (P'P)^{-1} L] + (\hat{\beta}_3^2 - \frac{R_o^2}{n-3} c_{33}) \hat{V}(s_x^2) \qquad (4.6)$$

is an unbiased estimator of $V(\bar{y}_p)$ determined by the observed sample values $y_1,\ldots,y_n$ and $x_1,\ldots,x_n$.

## REFERENCES


1. Cochran, W.G. (1963) Sampling Techniques, 2nd edition, John Wiley and Sons, New York.

2. Hansen, M.H., Hururitz, W.M. and Hadow, W.G. (1953) Sample Survey Methods and Theory, Vol. II, John Wiley, New York.

3. Kendall, M.G. and Stuart, A. (1969) "The Advanced Theory of Statistics," Vol I, 3rd edition, Hafner Publishing Company, New York.

4. Rao, C.R. (1973) "Linear Statistical Inference and Its Applications," 2nd edition, John Wiley & Sons, Inc., New York.

5. Sukhatme, P.V. and Sukhatme, B.V. (1970) "Sampling Theory of Surveys With Applications", Iowa State University Press, Ames, Iowa.


*This problem was initially suggested by the late H.O. Hartly.