Journal of Applied Sciences

ISSN 1812-5654

Journal of Applied Sciences 8 (8): 1597-1599, 2008
ISSN 1812-5654

Polynomial (Non Linear) Regression Method for Improved Estimation Based on Sampling

[1]Emmanuel John Ekpenyong, [1]Mfon Ime Okonnah and [2]Eno Donatus John
[1]Department of Statistics, Akwa Ibom State Polytechnic, Ikot Osurua, Nigeria
[2]Mathematics Unit, Department of General Studies, Akwa Ibom State Polytechnic, Ikot Osurua, Nigeria

Abstract: We seek to proffer estimates of the parameters of the dependent variable having a polynomial relationship with the independent variable, specifically that of order two. The variances of the mean and total estimates of the dependent variable are derived. The efficiency and precision of the method are shown by comparing its estimates with those of linear regression and elemental sampling methods through the use of sampled data by simple random sampling without replacement from simulated data.

Key words: Polynomial regression, auxiliary variables, linear regression, elemental sampling, quadratic model, estimation

INTRODUCTION

Regression analysis is often used in the analysis of survey data in situations where complex sampling designs are employed. Okafor (2002) highlighted that regression estimation method uses auxiliary information to improve the estimates of the population parameters such as the mean and total. He further noted that the regression estimation is used to estimate the population mean when the regression line of y on x does not pass through the origin but makes an intercept along the y-axis.

Matloff (1981) obtained the estimate of the unconditional mean μ of the variable Y from a random sample of a population having a distribution F_{xy}. He compared this estimate with that obtained from the linear regression of Y, the dependent variable on X, the auxiliary (independent) variable through averaging of the estimated regression function values at the sample points. He found out that the latter estimator offered substantial improvement over the sample mean of Y than the former.

In another development, Jewell and Queensberry (1986) considered an iterative regression method from data collected from stratified samples using a design variable, which is correlated with the dependent variable but is not included as an independent variable. They showed that the method gave a general superior estimate of the mean in terms of efficiency.

In all these research, the linear regression of a variable of interest Y (dependent variable) on the auxiliary X (independent variable) was considered for the estimation of population parameters. This study therefore seeks to consider a case where the variable of interest Y has a polynomial (non-linear) relationship with the auxiliary variable X, which is the independent variable. The study considers mainly the polynomial model of order two given by

$$Y = \beta_0+\beta_1X+\beta_2X^2+E$$

Ratkowsky (1983) indicated that numerous nonlinear models such as Weilbull-type model had been used to model sigmoidal growth curves widespread in biology, agriculture, engineering and economics. For a standard non-linear regression analysis, a correct regression model is often assumed to exist. This is a model whose algebraic sum of the error is zero. Statistical techniques for estimating the model parameters have been developed under the assumption of independent observations of Draper and Smith (1981). Asymptotic properties of least squares estimators have been discussed extensively in the literature such as Gallant (1987) and Wu (1981).

Bunke (1993) showed a high order asymptotic equivalence between extended Jacknife and asymptotic estimates in nonlinear regression model. Hung (1985) considered regression estimation with transformed auxiliary variates.

MATERIALS AND METHODS

A simple random sample of size 100 is selected by simple random sampling without replacement (srswor) from the generated sets of dependent (Y) and independent (X) variables.

Corresponding Author: Emmanuel John Ekpenyong, Department of Statistics, Akwa Ibom State Polytechnic, Ikot Osurua, Nigeria

Derivation of estimators of the population parameters (mean and total) and their corresponding variances: Given the general quadratic equation.

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + E \quad (1)$$

(1) can be transformed into an intrinsically linear model as:

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + E \quad (2)$$

Where:
$Z = X^2$

therefore the unbiased estimator of the mean is given by:

$$\bar{y}_{pl} = \bar{y} - \hat{\beta}_1(\bar{x} - \bar{X}) - \hat{\beta}_2(\bar{z} - \bar{Z}) \quad (3)$$

its variance is given by:

$$V(\bar{y}_{pl}) = E\left[\bar{y} - \hat{\beta}_1(\bar{x} - \bar{X}) - \hat{\beta}_2(\bar{z} - \bar{Z})\right]^2 - E^2\left[\bar{y} - \hat{\beta}_1(\bar{x} - \bar{X}) - \hat{\beta}_2(\bar{z} - \bar{Z})\right]$$

but

$$\begin{aligned} &E\left[\bar{y} - \hat{\beta}_1(\bar{x} - \bar{X}) - \hat{\beta}_2(\bar{z} - \bar{Z})\right]^2 \\ &= E\left\{\left[\bar{y} - \hat{\beta}_1(\bar{x} - \bar{X}) - \hat{\beta}_2(\bar{z} - \bar{Z})\right]\left[\bar{y} - \hat{\beta}_2(\bar{x} - \bar{X}) - \hat{\beta}_2(\bar{z} - \bar{Z})\right]\right\} \end{aligned} \quad (4)$$

$$\begin{aligned} &E[\bar{y}^2 - \hat{\beta}_1\bar{y}(\bar{x} - \bar{X}) - \hat{\beta}_2\bar{y}(\bar{z} - \bar{Z}) - \hat{\beta}_1\bar{y}(\bar{x} - \bar{X}) + \hat{\beta}_2^2(\bar{x} - \bar{X})^2 \\ &+ \hat{\beta}_1\hat{\beta}_2(\bar{x} - \bar{X})(\bar{z} - \bar{Z}) - \hat{\beta}_1\bar{y}(\bar{x} - \bar{X}) + \hat{\beta}_1\hat{\beta}_2(\bar{x} - \bar{X})(\bar{z} - \bar{Z}) \\ &+ \hat{\beta}_2^2(\bar{x} - \bar{X})^2] \\ &= E[\bar{y}^2 - 2\hat{\beta}_1\bar{y}(\bar{x} - \bar{X}) - 2\hat{\beta}_2\bar{y}_2(\bar{x} - \bar{X}) + 2\hat{\beta}_1\hat{\beta}_2(\bar{x} - \bar{X}) \\ &(\bar{z} - \bar{Z}) - \hat{\beta}_1^2(\bar{x} - \bar{X})^2 + \hat{\beta}_2^2(\bar{z} - \bar{Z})^2] \end{aligned} \quad (5)$$

Also

$$\begin{aligned} &E^2[\bar{y} - \hat{\beta}_1(\bar{x} - \bar{X}) - \hat{\beta}_2(\bar{z} - \bar{Z})] \\ &= \{E[\bar{y} - \hat{\beta}_1(\bar{x} - \bar{X}) - \hat{\beta}_2(\bar{z} - \bar{Z})]\}^2 \\ &= [E(\bar{y}) - \hat{\beta}_1 E(\bar{x} - \bar{X}) - \hat{\beta}_2 E(\bar{z} - \bar{Z})]^2 \\ &= [E(\bar{y})]^2 - \hat{\beta}_1 E(\bar{y})E(\bar{x} - \bar{X}) - \hat{\beta}_2 E(\bar{y})E(\bar{z} - \bar{Z}) \\ &- \hat{\beta}_1 E(\bar{y})E(\bar{x} - \bar{X}) + \hat{\beta}_1^2 E^2(\bar{x} - \bar{X}) \\ &+ \hat{\beta}_1\hat{\beta}_2 E(\bar{x} - \bar{X})E(\bar{z} - \bar{Z}) - \hat{\beta}_2 E(\bar{z} - \bar{Z})E(\bar{y}) \\ &+ \hat{\beta}_1\hat{\beta}_2 E(\bar{x} - \bar{X})E(\bar{z} - \bar{Z}) + \hat{\beta}_2 E^2(\bar{z} - \bar{Z}) \end{aligned} \quad (6)$$

Therefore, $V(\bar{y}pl)$ = (5)−(6) which gives:

$$\begin{aligned} &= V(\bar{y}) - 2\hat{\beta}_1 \operatorname{cov}(\bar{y},\bar{x}) - 2\hat{\beta}_2 \operatorname{cov}(\bar{y},\bar{z}) + 2\hat{\beta}_1\hat{\beta}_2 \operatorname{cov}(\bar{x},\bar{z}) \\ &+ \hat{\beta}_1^2 V(\bar{x}) + \hat{\beta}_2^2 V(\bar{z}) \end{aligned} \quad (7)$$

$$\therefore V(\bar{y}pl) = \frac{1-f}{n}\left[S_y^2 - 2\hat{\beta}_1 S_{yx} - 2\hat{\beta}_2 S_{yz} + 2\hat{\beta}_1\hat{\beta}_2 S_{xz} + \hat{\beta}_1^2 S_x^2 + \hat{\beta}_2^2 S_z^2\right] \quad (8)$$

We need to estimate β_1 and β_2 such that $V(\bar{y}pl)$ is a minimum. By the method of ordinary least squares, we differentiate partially (7) with respect to β_1 and β_1 to obtain the following normal equations:

$$\hat{\beta}_2 \operatorname{cov}(\bar{x},\bar{z}) + \hat{\beta}_2 V(\bar{x}) = \operatorname{Cov}(\bar{y},\bar{x}) \quad (9)$$

$$\hat{\beta}_2 V(\bar{x}) + \hat{\beta}_1 \operatorname{Cov}(\bar{x},\bar{z}) = \operatorname{Cov}(\bar{y},\bar{z}) \quad (10)$$

Solving (9) and (10) simultaneously, we obtain

$$\hat{\beta}_2 = \frac{\operatorname{Cov}(\bar{y},\bar{x})\ \operatorname{Cov}(\bar{x},\bar{z}) - \operatorname{Cov}(\bar{y},\bar{z})V(\bar{x})}{[\operatorname{Cov}(\bar{x},\bar{z})]^2 - V(\bar{x})V(\bar{z})} \quad (11)$$

$$\hat{\beta}_1 = \frac{\operatorname{Cov}(\bar{y},\bar{z})\ \operatorname{Cov}(\bar{x},\bar{z}) - \operatorname{Cov}(\bar{y},\bar{x})\ V(\bar{z})}{[\operatorname{Cov}(\bar{x},\bar{z})]^2 - V(\bar{x})V(\bar{z})} \quad (12)$$

or

$$\hat{\beta}_2 = \frac{s_{yx}s_{xz} - s_{yz}s_x^2}{s_{xz}^2 - s_x^2 s_z^2} \quad (13)$$

$$\hat{\beta}_1 = \frac{s_{yz}s_{xz} - s_{yx}s_z^2}{s_{xz}^2 - s_x^2 s_z^2} \quad (14)$$

and the total estimate and its variance are given as:

$$ypl = N\bar{y}_{pl} \quad (15)$$

$$V(\bar{y}_{pl}) = N^2 V(\bar{y}_{pl}) \quad (16)$$

To determine the efficiency and precision of the above method, the estimates will be compared with the estimates from Linear regression and elemental sampling method. Any method with the least variance gives the method with a more efficient and precise estimate.

We note here that s_{xy}, s_{zy}, s_{xz} are estimators of the population covariances S_{xy}, S_{zy} and S_{xy}, respectively; while variances s_x^2, s_y^2, s_z^2 are unbiased estimators of the population S_x^2, S_y^2, S_z^2, respectively.

DATA ANALYSIS

s_{xy} = 3271.758, s_{zy} = 37433.466, s_{zx} = 16.208
s^2_x = 1.4462, s^2_y = 8046950.712, s^2_z = 185.760

From Eq. 13 and 14.

$$\hat{\beta}_1 = 174.70,\ \hat{\beta}_2 = 186.27$$

$$\begin{aligned} \text{and using Eq. (3)}: \bar{y}_{pl} &= 5890.43 - 174.70(5.208 - 5.202) \\ &- 186.27(28.555 - 28.313) \\ &= 5844.30 \end{aligned}$$

Also the variance of $\bar{y}_{pl}$ is obtained by Eq. 8.

$$\therefore\ V(\bar{y}_{pl}) = 0.005[8046950.712 - 2(174.70)(3271.758) - 2(186.27)(37433.47) + 2(174.7)(186.27)(16.208) + (174.7)^2(1.4462) + (186.27)^2(185.76)]$$
$$= 8046950.712 - 13945463 + 1054861 + 44138.154 + 6445224.3 = 2512.795$$

The means and their variances in Linear regression and elemental sampling method are:

$$\bar{y}lr = \bar{y} - \hat{\beta}(\bar{x} - \bar{X}) = 5890.43 - 2262.42(5.208 - 5.202) = 5876.86$$

$$V(\bar{y}lr) = \frac{1-f}{n}\left[s_y^2 + \hat{\beta}^2 s_x^2 - 2\hat{\beta}s_{xy}\right] = 3226.0425$$

For elemental sampling: $\bar{y} = 5890.43$

$$V(\bar{y}) = \frac{1-f}{n}s_y^2 = 0.005 \times 8046950.712 = 40234.754$$

Their total estimates are given as follows:

$$y_{pl} = N\bar{y}_{pl} = 200 \times 5844.3 = 1168860,\ V(\bar{y}pl) = N^2V(\bar{y}lr) = 100511800$$
$$y_{lr} = N\bar{y}_{lr} = 200 \times 5876.86 = 1175372,\ V(ylr) = N^2V(\bar{y}lr) = 129041700$$
$$\bar{y} = N\bar{y} = 200 \times 5890.43 = 1178086,\ V(y) = N^2V(\bar{y}) = 1609390160$$

RESULTS AND DISCUSSION

From Table 1, the polynomial (quadratic) regression method provides estimates with the least variances and this is followed by those of linear regression method. Therefore, the polynomial regression method gives the most efficient estimates of the population parameters (Mean and Total).

The efficiency and precision of using Regression Estimation Method depend on the degree of relationship between the auxiliary variable and the variable of interest (independent variable). It also depends on their nature or pattern of relationship and the sample size. Suppose that the appropriate relationship between the auxiliary and independent variable is a polynomial one and one mispecifies the model and estimates the population parameters using the Linear regression estimation method, it will lead to an inefficient or less efficient parameters estimation. A similar thing will be obtained when the degree of relationship is low. It is also noteworthy that the appropriate order of polynomial is imperative for a polynomial relationship.

This research at a variable that has a polynomial relationship with an auxiliary variable will have its parameters well estimated using the Polynomial regression method which will give more efficient and precise estimates.

Table 1: Summary of estimates of means, totals and their variances

	Methods		
Estimates	Polynomial regression	Linear regression	Elemental sampling
Mean($\bar{y}$)	5844.300	5876.8600	5890.430
V($\bar{y}$)	2512.795	3226.0405	40234.754
Total(y)	1168860.000	1175372.0000	1178086.000
V(y)	100511800.000	129041700.0000	1609390160.000

CONCLUSION

Polynomial regression estimation method provides a way of making good, efficient and precise estimates of characteristics of variables that have polynomial relationships with their auxiliary variables. It is therefore necessary for one to understand the nature and pattern of relationship between a variable of interest and the independent variable before choosing a method for estimation of parameters.

REFERENCES

Bunke, O., 1993. Extended jackknife estimates in linear or nonlinear regression. JTAS., 15 (1): 47-61.

Draper, N.R. and H. Smith, 1981. Applied Regression Anal. John Wiley, New York.

Gallant, A.R., 1987. Non Linear Statistical Models. John Wiley, New York.

Hung, H.M., 1985. Regression estimation with transformed auxiliary variates. Stat. Probability Lett., 5: 239-243.

Jewell, N.P. and C.P. Queensberry, Jr., 1986. Regression analysis based on stratified samples. Brometrika, 73 (3): 605-661.

Matloff, N.S., 1981. Use of regression functions for improved estimation of means. Biometrika, 68 (3): 685-689.

Okafor, F.C., 2002. Sample Survey Theory With Appls. Afro-Orbis Publishers, Nsukka, Nigeria.

Ratkowsky, D.A., 1983. Nonlinear Regression Modeling. Marcel Dekker, New York.

Wu, C.F., 1981. Asymptotic theory of nonlinear least squares estimation. Ann. Stat., 9: 501-513.