905.2

PACIFIC SOUTHWEST
Forest and Range
Experiment Station

FOREST SERVICE
U.S.DEPARTMENT OF AGRICULTURE
P.O. BOX 245, BERKELEY, CALIFORNIA 94701

REGRESSION SAMPLING:

some results for resource managers and researchers

William G. O'Regan   Robert W. Boyd

USDA Forest Service
Research Note PSW-286
1974

Abstract: Regression sampling is widely used in natural resources management and research to estimate quantities of resources per unit area. This note brings together results found in the statistical literature in the application of this sampling technique. Conditional and unconditional estimators are listed and for each estimator, exact variances and unbiased estimators for the variances are offered.

Oxford: 905.2–015.5.
Retrieval Terms: regression sampling; statistical techniques.

Workers in natural resources employ a number of statistical techniques to increase the effectiveness of their sampling activities. These techniques include cluster, stratified, two-stage stratified, and regression sampling. These methods are used, with varying degrees of effectiveness, to estimate quantities of resources per unit area.

Regression sampling is one of the more widely used methods. Users attempt to take advantage of the relationship between the variable of interest (Y), which is usually expensive to measure, and a related variable (X), which is usually less costly to measure.

This note collects, organizes, and reports some results of interest to those using regression sampling. The results can be found scattered in the literature[1] and are exact within the confines of the stated assumptions. Listed here are conditional and unconditional estimators and for each estimator, exact variances and unbiased estimators for the variances.

POPULATION OF INTEREST

We postulate the following relationships in the population of interest:

$$[Y|X_i] = \theta_y + \beta(X_i - \theta_x) + \epsilon_i$$
$$\epsilon \rightarrow N(0, \sigma_\epsilon{}^2)$$
$$X \rightarrow N(\theta_x, \sigma_x{}^2)$$

It follows that

$$Y \rightarrow N(\theta_y, \sigma_\epsilon^2 + \beta^2 \sigma_x^2)$$

The following sampling procedure is specified:

1. Take a random sample size $\underline{n}$, from a large population.
2. Measure X and Y.

3. Calculate $\overline{Y}, \hat{\beta}$, and $\overline{X}$, estimates of θ_y, β, and θ_x.

$$\overline{Y} = \Sigma Y/n$$
$$\overline{X} = \Sigma X/n$$
$$\hat{\beta} = \Sigma xy/\Sigma x^2$$

in which

$$x = X - \overline{X}$$
$$y = Y - \overline{Y}$$

4. For a value of X, say X*, use

$$Y^* = \overline{Y} + \hat{\beta}(X^* - \overline{X})$$

as the estimated value of Y.

5. When X* is random, it is considered to be the mean of N independent, random observations on X, i.e.,

$$X^* = \Sigma X/N$$

6. When X* is not random it is a fixed value of X. In some situations,

$$X^* = E(X) = \theta_x.$$

Questions about the expected value and variance of Y* can be answered only within the context of the sampling distribution of Y*. This distribution depends upon the nature of the "repeated sampling" that gives rise to the sampling distribution.

There are four possibilities, or cases, with two additional special cases:

(1) The n X's in the first sample (a set, S_1) and X* are considered fixed in the repeated samples; i.e., only the n Y's vary.

(1s) Same as 1, except that $X^*=\theta_x$.

(2) S_1 and the Y's vary, while X* is fixed. In which case

$$\overline{X} \rightarrow N(\theta_x, \sigma_x^2/n)$$

(2s) Same as 2, except that $X^*=\theta_x$.

(3) S_1 is fixed, while the n Y's and X* are random, and

$$X^* \rightarrow N(\theta_x, \sigma_x^2/N)$$

(4) S_1, X*, and the n Y's all vary in repeated samples.

Expected values, variances, and unbiased estimates of the variance of Y* for each of these specifications are listed in *table 1*.

CHOICE OF MODELS

The choice of models is a subject matter problem—not a statistical problem. How the values, variances, and estimated variances listed in *table 1* are used will depend on which specification describes the subject matter problem. Furthermore, the user should remember the normality and independence assumptions, that n and N are fixed sample sizes, that "given S_1, X*" means that only the n observations on Y are conceived of as varying in repeated samples, and that "given X*" means that the n observations of paired X,Y are variable. "Given X*" indicates that the estimate is for some arbitrary value of X=X* (sometimes $X^*=\theta_x$; that is, the mean of the distribution of X's is known). Obviously, "given S_1" implies that the n observations on Y and the N observations on X vary from sample to sample. The applicability of these specifications depends upon the problem at hand *(table 1)*.

The validity of the entries in columns 3, 4, and 5 of *table 1* can be demonstrated:

The relationships

$$E(X) = E[E(X|Y)]$$
$$V(X) = E[X-E(X)]^2 = E(X^2) - [E(X)]^2$$

in which X and Y are random variables are used throughout these demonstrations.

We know

$$E(1/\chi_p^2) = 1/(p-2) \tag{1}$$

and

$$E(F_{p,q}) = q/(q-2). \tag{2}$$

From the specifications,

$$\overline{Y} = \Sigma Y/n = \theta_y + \beta(\overline{X}-\theta_x) + \overline{\epsilon}. \tag{3}$$

All variables in this development are to be assumed subscripted by i. All summations are over i (i=1, 2, . . ., n). Therefore, we omit all limits of summation.

Then,

$$\overline{\epsilon} = \Sigma\epsilon/n$$
$$E(\overline{\epsilon}) = 0 \tag{4}$$

$$E(\overline{\epsilon}^2) = \sigma_\epsilon^2/n \tag{5}$$

$$\hat{\beta} = \Sigma xY/\Sigma x^2 = \Sigma x(\theta_y + \beta(X-\theta_x) + \epsilon)/\Sigma x^2 = \beta + \overline{\epsilon}_w \tag{6}$$

$$\overline{\epsilon}_w = \Sigma x\epsilon/\Sigma x^2$$
$$E(\overline{\epsilon}_w) = 0 \tag{7}$$

Table 1–*Expected values, variances, and estimated variances for Y* (the estimated value in regression sampling), for different specifications*[1]

Cases	Specifications	E(Y*)	V(Y*)	$\hat{V}$(Y*)
(1)	S_1, X* fixed	$\theta_y + \beta(X^*-\theta_x)$	$\sigma_\epsilon^2[1/n + (X^*-\overline{X})^2/\Sigma x^2]$	$\hat{\sigma}_\epsilon^2[1/n + (X^*-\overline{X})^2/\Sigma x^2]$
(1s)	S_1 fixed; $X^* = \theta_x$	θ_y	$\sigma_\epsilon^2[1/n + (\theta_x-\overline{X})^2/\Sigma x^2]$	$\hat{\sigma}_\epsilon^2[1/n + (\theta_x-\overline{X})^2/\Sigma x^2]$
(2)	S_1 random, X* fixed	$\theta_y + \beta(X^*-\theta_x)$	$[\sigma_\epsilon^2/(n-3)]\ [(n-2)/n$ $+ (X^*-\theta_x)^2/\sigma_x^2]$	$\hat{\sigma}_\epsilon^2[1/n + (X^*-\overline{X})^2/\Sigma x^2]$
(2s)	S_1 random, $X^* = \theta_x$	θ_y	$(\sigma_\epsilon^2/n)\big((n-2)/(n-3)\big)$	$(\hat{\sigma}_\epsilon^2/n)\big((n-2)/(n-3)\big)$
(3)	S_1 fixed, X* random	θ_y	$\sigma_\epsilon^2[1/n + \big((\overline{X}-\theta_x)^2 + (\sigma_x^2/N)\big)/\Sigma x^2]$ $+ \beta^2\sigma_x^2/N$	$\hat{\sigma}_\epsilon^2[1/n + (X^*-\overline{X})^2/\Sigma x^2$ $-(\hat{\sigma}_x^2/N)/\Sigma x^2] + \hat{\beta}^2\hat{\sigma}_x^2/N$
(4)	S_1, X* random	θ_y	$(\sigma_\epsilon^2/n)\big((n-2)/(n-3)\big)$ $+ (\sigma_\epsilon^2/N)/(n-3) + \beta^2\sigma_x^2/N$	$(\hat{\sigma}_\epsilon^2/n)\big((n-2)/(n-3)\big)$ $+ \hat{\beta}^2\hat{\sigma}_x^2/N$

[1] The n observations on Y are always variable in "repeated" samples.

$$E(\bar{\epsilon}_w^2 | S_1) = \sigma_\epsilon^2 / \Sigma x^2 \tag{8}$$

$$E(\bar{\epsilon}_w^2) = E(\sigma_\epsilon^2/\Sigma x^2) = E[(\sigma_\epsilon^2/\sigma_x^2)/(\Sigma x^2/\sigma_x^2)] = (\sigma_\epsilon^2/\sigma_x^2)E(1/\chi^2_{n-1}) = (\sigma_\epsilon^2/\sigma_x^2)/(n-3) \tag{9}$$

$$E(\bar{\epsilon}\,\bar{\epsilon}_w) = 0 \tag{10}$$

$$E[(\bar{X}-\theta_x)/\Sigma x^2] = 0 \tag{11}$$

$$\begin{aligned} E[(\bar{X}-\theta_x)^2/\Sigma x^2] &= E\left[\left(\frac{(\bar{X}-\theta_x)^2}{\sigma_x^2/n}\cdot \sigma_x^2/n\right) \Big/ \left(\frac{\Sigma x^2}{\sigma_x^2(n-1)} \cdot \sigma_x^2(n-1)\right)\right] \\ &= \frac{\sigma_x^2/n}{\sigma_x^2(n-1)}\, E\left[\frac{(\bar{X}-\theta_x)^2}{\sigma_x^2/n} \Big/ \frac{\Sigma x^2}{\sigma_x^2(n-1)}\right] \\ &= \frac{1}{n(n-1)}\, E[F_{1,n-1}] \\ &= \frac{n-1}{n(n-1)(n-3)} \\ &= \frac{1}{n(n-3)} \end{aligned} \tag{12}$$

The estimated value (Y^*) for a value of $X = X^*$ is given by

$$Y^* = \bar{Y} + \hat{\beta}(X^* - \bar{X}) \tag{13}$$

$$\begin{aligned} &= \theta_y + \beta(\bar{X}-\theta_x) + \bar{\epsilon} + (\beta + \bar{\epsilon}_w)(X^* - \bar{X}) \\ &= \theta_y + \beta(X^*-\theta_x) + \bar{\epsilon}_w(X^*-\theta_x) - \bar{\epsilon}_w(\bar{X}-\theta_x) + \bar{\epsilon} \end{aligned} \tag{14}$$

$$E(Y^* | S_1, X^*) = \theta_y + \beta(X^* - \theta_x) \tag{15}$$

$$E(Y^* | S_1, X^* = \theta_x) = \theta_y \tag{16}$$

$$E(Y^* | S_1) = \theta_y \tag{17}$$

$$E(Y^* | X^*) = \theta_y + \beta(X^* - \theta_x) \tag{18}$$

$$E(Y^* | X^* = \theta_x) = \theta_y \tag{19}$$

$$E(Y^*) = \theta_y \tag{20}$$

$$Y^{*2} = [\theta_y + \beta(X^*-\theta_x) + \bar{\epsilon}_w(X^*-\theta_x) - \bar{\epsilon}_w(\bar{X}-\theta_x) + \bar{\epsilon}]^2$$

$$Y^{*2} = A + B + C + D + F$$

in which

$$A = \theta_y^2 + \beta^2(X^*-\theta_x)^2 + \bar{\epsilon}_w(X^*-\theta_x)^2 + \bar{\epsilon}_w^2(\bar{X}-\theta_x)^2 + \bar{\epsilon}^2$$
$$B = 2\theta_y[\beta(X^*-\theta_x) + \bar{\epsilon}_w(X^*-\theta_x) - \bar{\epsilon}_w(\bar{X}-\theta_x) + \bar{\epsilon}]$$
$$C = 2\beta[\bar{\epsilon}_w(X^*-\theta_x)^2 - \bar{\epsilon}_w(X^*-\theta_x)(\bar{X}-\theta_x) + (X^*-\theta_x)\bar{\epsilon}]$$
$$D = 2[-\bar{\epsilon}_w^2(X^*-\theta_x)(\bar{X}-\theta_x) + \bar{\epsilon}_w\bar{\epsilon}(X^*-\theta_x)]$$
$$F = -2\bar{\epsilon}_w\bar{\epsilon}(\bar{X}-\theta_x)$$

$$E(A|S_1, X^*) = \theta_y^2 + \beta^2(X^*-\theta_x)^2 + (X^*-\theta_x)^2\,\sigma_\epsilon^2/\Sigma x^2 + (\bar{X}-\theta_x)^2\,(\sigma_\epsilon^2/\Sigma x^2) + \sigma_\epsilon^2/n \qquad (21)$$

$$E(A|S_1, X^*=\theta_x) = \theta_y^2 + (\bar{X}-\theta_x)^2(\sigma_\epsilon^2/\Sigma x^2) + \sigma_\epsilon^2/n \qquad (22)$$

$$E(A|S_1) = \theta_y^2 + \beta^2\sigma_x^2/N + (\sigma_\epsilon^2/\Sigma x^2)(\sigma_x^2/N) + (\bar{X}-\theta_x)^2(\sigma_\epsilon^2/\Sigma x^2) + \sigma_\epsilon^2/n \qquad (23)$$

$$\begin{aligned}E(A|X^*) &= \theta_y^2 + \beta^2(X^*-\theta_x)^2 + (X^*-\theta_x)^2(\sigma_\epsilon^2/\sigma_x^2)E[1/\chi^2_{n-1}] \\ &\quad + \left(\sigma_\epsilon^2/(n(n-1))\right)\ E[F_{1,n-1}] + \sigma_\epsilon^2/n \\ &= \theta_y^2 + \beta^2(X^*-\theta_x)^2 + \left((X^*-\theta_x)^2/(n-3)\right)(\sigma_\epsilon^2/\sigma_x^2) + \sigma_\epsilon^2/n(n-3) + \sigma_\epsilon^2/n \\ &= \theta_y^2 + \beta^2(X^*-\theta_x)^2 + (X^*-\theta_x)^2(\sigma_\epsilon^2/\sigma_x^2)/(n-3) \\ &\quad + \left((\sigma_\epsilon^2/n)\right)\left((n-2)/(n-3)\right)\end{aligned} \qquad (24)$$

$$E(A|X^*=\theta_x) = \theta_y^2 + \left(\sigma_\epsilon^2/n\right)\left((n-2)/(n-3)\right) \qquad (25)$$

$$\begin{aligned}E(A) &= \theta_y^2 + \beta^2\sigma_x^2/N + (\sigma_x^2\sigma_\epsilon^2/N)E[1/\Sigma x^2] + \sigma_\epsilon^2 E[(\bar{X}-\theta_x)^2/\Sigma x^2] + \sigma_\epsilon^2/n \\ &= \theta_y^2 + \beta^2\sigma_x^2/N + \sigma_\epsilon^2/N(n-3) + \left(\sigma_\epsilon^2/n\right)\left((n-2)/(n-3)\right)\end{aligned} \qquad (26)$$

$$E(B|S_1, X^*) = E(B|X^*) = 2\theta_y\beta(X^*-\theta_x) \qquad (27)$$

$$E(B|S_1, X^*=\theta_x) = E(B|S_1) = E(B|X^*=\theta_x) = E(B) = 0 \qquad (28)$$

$$\begin{aligned}E(C|S_1, X^*) &= E(C|S_1, X^*=\theta_x) = E(C|S_1) = E(C|X^*) \\ &= E(C|X^*=\theta_x) = E(C) = 0\end{aligned} \qquad (29)$$

$$E(D|S_1, X^*) = -2\sigma_\epsilon^2(X^*-\theta_x)(\bar{X}-\theta_x)/\Sigma x^2 \qquad (30)$$

$$E(D|S_1, X^*=\theta_x) = E(D|S_1) = E(D|X^*) = E(D|X^*=\theta_x) = E(D) = 0 \qquad (31)$$

$$E(F|S_1, X^*) = E(F|S_1) = E(F|X^*) = E(F|X^*=\theta_x) = E(F) = 0 \qquad (32)$$

From 21, 27, 29, 30, 32, and 15

$$V(Y^*|S_1, X^*) = \sigma_\epsilon^2/n + (X^*-\bar{X})^2 (\sigma_\epsilon^2/\Sigma x^2) \tag{33}$$

From 22, 28, 29, 31, 32, and 16

$$V(Y^*|S_1, X^*=\theta_x) = \sigma_\epsilon^2/n + (\bar{X}-\theta_x)^2 (\sigma_\epsilon^2/\Sigma x^2) \tag{34}$$

From 23, 28, 29, 31, 32, and 17

$$V(Y^*|S_1) = \sigma_\epsilon^2/n + \left(\sigma_x^2/N + (\bar{X}-\theta_x)^2\right) (\sigma_\epsilon^2/\Sigma x^2) + \beta^2 \sigma_x^2/N \tag{35}$$

From 24, 27, 29, 31, 32, and 18

$$V(Y^*|X^*) = (\sigma_\epsilon^2/n)\left((n-2)/(n-3)\right) + \left(\sigma_\epsilon^2/(n-3)\right)\left((X^*-\theta_x)^2/\sigma_x^2\right) \tag{36}$$

From 25, 28, 29, 31, 32, and 19

$$V(Y^*|X^*=\theta_x) = (\sigma_\epsilon^2/n)\left((n-2)/(n-3)\right) \tag{37}$$

From 26, 28, 29, 31, 32, and 20

$$V(Y^*) = (\sigma_\epsilon^2/n)\left((n-2)/(n-3)\right) + (\sigma_\epsilon^2/N)/(n-3) + \beta^2 \sigma_x^2/N \tag{38}$$

If we define

$$\hat{\sigma}_\epsilon^2 = \frac{\Sigma[Y-\bar{Y}-\hat{\beta}(X-\bar{X})]^2}{n-2}$$

the unbiasedness of the variance estimators for cases (1), (1s) and (2s) is clear.

Substituting $(X^*-\theta_x)-(\bar{X}-\theta_x)$ for $(X^*-\bar{X})$, the expression for the estimated variance for case (2) can be written

$$\begin{aligned}
\hat{V}(Y^*|X^*) &= \hat{\sigma}_\epsilon^2/n + \hat{\sigma}_\epsilon^2[(X^*-\theta_x)^2/\Sigma x^2 + 2(X^*-\theta_x)(\bar{X}-\theta_x)/\Sigma x^2 \\
&\quad + (\bar{X}-\theta_x)^2/\Sigma x^2] \\
E\hat{V}(Y^*|X^*) &= \sigma_\epsilon^2/n + \sigma_\epsilon^2[(X^*-\theta_x)^2 E(1/\Sigma x^2) - 2(X^*-\theta_x)E\left((\bar{X}-\theta_x)/\Sigma x^2\right) \\
&\quad + E\left((\bar{X}-\theta_x)^2/\Sigma x^2\right)] \\
&= \sigma_\epsilon^2/n + \sigma_\epsilon{}^2[(X^*-\theta_x)^2/(\sigma_x^2(n-3)) - 0 + 1/(n(n-3))] \\
&= (\sigma_\epsilon^2/(n-3))\left((n-2)/n + (X^*-\theta_x)^2/\sigma_x^2\right)
\end{aligned}$$

For case (3) we have

$$\begin{aligned}
E(\hat{\beta}^2\hat{\sigma}_x^2/N|S_1, X^*) &= (\hat{\sigma}_x^2/N)(\beta^2 + \sigma_\epsilon^2/\Sigma x^2) \\
E(\hat{\beta}^2\hat{\sigma}_x^2/N|S_1) &= (\sigma_x^2/N)(\beta^2 + \sigma_\epsilon^2/\Sigma x^2)
\end{aligned} \tag{39}$$

$$\begin{aligned}
E[(\hat{\sigma}_\epsilon^2)&\left(1/n + (\bar{X}-X^*)^2/\Sigma x^2 - (\hat{\sigma}_x^2/N)/\Sigma x^2\right)|S_1, X^*] \\
&= \sigma_\epsilon^2\left(1/n + (\bar{X}-X^*)^2/\Sigma x^2 - (\hat{\sigma}_x^2/N)/\Sigma x^2\right) \\
E[(\bar{X}-X^*)^2|S_1] &= E[\left((\bar{X}-\theta_x) + (X^*-\theta_x)\right)^2|S_1] \\
&= E[\left((\bar{X}-\theta_x)^2 + 2(\bar{X}-\theta_x)(X^*-\theta_x) + (X^*-\theta_x)^2\right)|S_1] \\
&= (\bar{X}-\theta_x)^2 + 0 + \sigma_x^2/N
\end{aligned} \tag{40}$$

$$E[(\hat{\sigma}_\epsilon^2/\Sigma x^2)(\hat{\sigma}_x^2/N)\mid S_1, X^*] = (\sigma_\epsilon^2/\Sigma x^2)(\hat{\sigma}_x^2/N)$$

$$E[(\hat{\sigma}_\epsilon^2/\Sigma x^2)(\hat{\sigma}_x^2/N)\mid S_1] = (\sigma_\epsilon^2/\Sigma x^2)(\sigma_x^2/N) \tag{41}$$

Using 39, 40, and 41, we have

$$E[\hat{\sigma}_\epsilon^2\ \big(1/n + (\bar{X}-X^*)^2/\Sigma x^2 - (\hat{\sigma}_x^2/N)/\Sigma x^2\big) + \hat{\beta}^2\hat{\sigma}_x^2/N \mid S_1]$$

$$= \sigma_\epsilon^2\ \big(1/n + (\bar{X}-\theta_x)^2/\Sigma x^2 + (\sigma_x^2/N)/\Sigma x^2 - (\sigma_x^2/N)/\Sigma x^2\big) + \beta^2\sigma_x^2/N$$

$$+ (\sigma_\epsilon^2/\Sigma x^2)(\sigma_x^2/N)$$

$$= \sigma_\epsilon^2\ \big(1/n + \big((\bar{X}-\theta_x)^2 + \sigma_x^2/N\big)/\Sigma x^2\big) + \beta^2\sigma_x^2/N$$

For case (4) we have

$$E(\hat{\beta}^2\hat{\sigma}_x^2/N \mid S_1, X^*) = (\hat{\sigma}_x^2/N)(\beta^2 + \sigma_\epsilon^2/\Sigma x^2)$$

$$E\big[(\hat{\sigma}_x^2/N)(\beta^2 + \sigma_\epsilon^2/\Sigma x^2)\mid S_1\big] = (\sigma_x^2/N)(\beta^2 + \sigma_\epsilon^2/\Sigma x^2)$$

$$= \beta^2\sigma_x^2/N + (\sigma_\epsilon^2/N)\big(1/(\Sigma x^2/\sigma_x^2)\big)$$

$$E\big[1/(\Sigma x^2/\sigma_x^2)\big] = E(1/\chi_{n-1}^2) = 1/(n-3)$$

and

$$E(\hat{\beta}^2\hat{\sigma}^2/N) = \beta^2\sigma_x^2/N + (\sigma_\epsilon^2/N)/(n-3)$$

which indicates that the variance estimator for case (4) is unbiased.

NOTE

[1]For examples, see Cochran, William G. *Sampling techniques.* Ed. 2. New York: John Wiley and Sons, Inc. p. 189-205. 1963; and Cunia, T. *Some theory on reliability of volume estimates in a forest inventory sample.* For. Sci. 11: 115-128. 1965.

The Authors

WILLIAM G. O'REGAN is chief of the Station's biometrics branch, and also a lecturer in the School of Forestry and Conservation, University of California, Berkeley. He earned a B.S. degree (1949) and a doctorate (1962) in agricultural economics at the University of California, Berkeley. **ROBERT W. BOYD** was formerly a statistician in the Station's biometrics branch. He is a 1972 graduate, in statistics, of the University of California, Berkeley.

The Forest Service of the U.S. Department of Agriculture

. . . Conducts forest and range research at more than 75 locations from Puerto Rico to Alaska and Hawaii.

. . . Participates with all State forestry agencies in cooperative programs to protect and improve the Nation's 395 million acres of State, local, and private forest lands.

. . . Manages and protects the 187-million-acre National Forest System for sustained yield of its many products and services.

The Pacific Southwest Forest and Range Experiment Station

represents the research branch of the Forest Service in California and Hawaii.

G. P. O. 794–474/4626