Conference on Applied Statistics in Agriculture 1992 - 4th Annual Conference Proceedings

ESTIMATING VARIANCE FUNCTIONS FOR WEIGHTED LINEAR REGRESSION

Michael S. Williams

Hans T. Schreuder

Timothy G. Gregoire

William A. Bechtold

See next page for additional authors

Recommended Citation

Williams, Michael S.; Schreuder, Hans T.; Gregoire, Timothy G.; and Bechtold, William A. (1992). "ESTIMATING VARIANCE FUNCTIONS FOR WEIGHTED LINEAR REGRESSION," *Conference on Applied Statistics in Agriculture*. https://doi.org/10.4148/2475-7772.1401

Author Information

Michael S. Williams, Hans T. Schreuder, Timothy G. Gregoire, and William A. Bechtold

ESTIMATING VARIANCE FUNCTIONS FOR WEIGHTED LINEAR REGRESSION

Michael S. Williams
Hans T. Schreuder
Multiresource Inventory Techniques
Rocky Mountain Forest and Range Experiment Station
USDA Forest Service
Fort Collins, Colorado 80526-2098, U.S.A.

Timothy G. Gregoire
Department of Forestry
Virginia Polytechnic Institute and State University
Blacksburg, Virginia 24061-0324, U.S.A.

William A. Bechtold
Research Forester
Southeastern Forest Experiment Station
Forest Inventory and Analysis Project
USDA Forest Service
Asheville, North Carolina 28802-2680, U.S.A.

ABSTRACT

For linear models with heterogeneous error structure, four variance function models are examined for predicting the error structure in two loblolly pine data sets and one white oak data set. An index of fit and a simulation study were used to determine which models were best. The size of coefficients for linear and higher order terms varied drastically across different data sets, thus it is not desirable to recommend a general model containing both linear and higher order terms. The unspecified exponent model $\sigma^2 v_i = \sigma^2 (D_i^2 H_i)^{k_1}$ is recommended for all data sets considered. The k_1 values ranged from 1.8 to 2.1. We recommend $k_1 = 2.0$ for simplicity.

1. INTRODUCTION AND LITERATURE REVIEW

Weighted linear regression estimation is widely used in forestry to estimate total volume (V) as a function of diameter (D) at breast height squared times total tree height (H). The model commonly used is

$$V_i = \alpha + \beta D_i^2 H_i + e_i \qquad [1]$$

where $D_i^2 H_i = x_i$, with $E[e_i e_j] = \sigma^2 x_i^{k_1}$ for $i = j$ $E[e_i e_j] = 0$ for $i \neq j$ and k_1 determines the rate of increase of the variance of the e_i's.

Cunia (1964) suggests $k_1 = 2$; McClure et al. (1983) suggest using $k_1 = 1.5$ for white oak and loblolly pine. Gregoire and Dyer (1989) also estimate the heterogeneity in volume equations. They assume model (1) and estimate k_1 in several ways for loblolly pine (n = 209) and red pine (n = 91) data sets. They obtain $k_1 = 1.70 - 1.84$ for the loblolly pine and $k_1 = 1.01 - 2.07$ for red pine using three estimation methods.

In the statistical literature Scott, Brewer, and Ho (1978) suggest that

$$E[e_i e_i] = k_2 x_i + k_3 x_i^2 \qquad [2]$$

might be generally suitable for data in many applications and Figure 3 in McClure et al.(1983) confirms this to some degree. Davidian and Carroll (1987) note that reliable variance estimates are important in yielding improved estimates of the regression coefficients and are also of interest in their own right. The authors consider the heteroscedastic error regression model $E[y_i] = f(x_i, \beta)$; Var $(y_i) = \sigma^2 g^2(z_i, \beta, \theta)$ where g expresses the heteroscedasticity, the z_i are a function of known factors (for example $D_i^2 H_i$), σ is an unknown scale parameter, and θ (r x 1) is a vector of unknown

parameters. No assumption about the distribution of the y_i other than the form of the first two moments is made. Davidian and Carroll (1987) state several definitive conclusions that apply only to symmetric error distributions but should not be ignored in practice.

1. Robustness is quite important in the efficiency of variance function estimation, probably even more so than in estimation of a mean function. The authors state, "For a standard contaminated normal model for which the best robust estimators have efficiency of 125% with respect to least squares, the absolute residual estimator of the variance function has efficiency of 200%".
2. Iteratively weighted least squares should be used so that the variance function estimate is based on generalized least squares residuals. The authors' experience has been that unweighted least squares residuals yield unreliable estimates of the variance function when the variances depend on the mean.
3. The precision of $\hat{\theta}$ is independent of σ.

Gregoire and Dyer (1989) discuss methods for fitting equation [1]. Their work indicates that maximum likelihood techniques produced superior results to methods based on squared residuals or transformed squared residuals.

In this paper, we examine potential improvements in the results of McClure et al. (1983) with equation [2] and other alternative variance functions using the same data sets but also a large additional data set to determine if a better error model exists. Maximum likelihood estimation procedures, recommended by Gregoire and Dyer (1989), are used to fit

$$V_i = \alpha + \beta x_i \text{ with } E[e_i e_j] = \sigma^2 v_i \text{ for } i = j \text{ where } v_i = f(x_i) \text{ and } E[e_i e_j] = 0 \text{ for } i \neq j. \quad [3]$$

The parameters to be estimated by maximum likelihood are the α, β, σ^2, and the k_i coefficients which determine the shape of the various different forms of the weighting function $f(x_i)$.

Use of such regression equations can result in significantly improved overall estimates of volume and also of the estimates of reliability of such estimates.

2. DATA DESCRIPTION

Southeastern Loblolly Pine and White Oak Populations

Individual sample trees used in this study were measured and tree volumes were computed using methods described by Cost (1978). There are 5,134 loblolly pine (Pinus taeda L.) trees and 1,484 white oak (Quercus alba L.) trees in these data sets (data was provided by Noel Cost, Southeastern Forest Experiment Station, Asheville, NC). Since 1963, the U.S. Forest Service Forest Inventory and Analysis (FIA unit) in the southeastern United States. has measured the volumes of individual standing trees on a subsample of all regular Forest Survey sample locations in Virginia, North Carolina, South Carolina, Georgia, and Florida. A supplemental sample of felled trees was also measured at hundreds of active logging operations distributed throughout the southeast. Trees were selected to ensure that an adequate sample of trees was taken from the range of tree diameters so that proportionally more large trees were selected for measurement than are found in the population. Trees were measured in a uniform manner by highly trained inventory specialists during the regular periodic inventories of the southeastern states. Special crews used marked sectional aluminum poles (McClure 1968) for length measurements and a McClure mirror caliper (McClure 1969) to measure upper-stem diameters in standing trees. All trees were measured as a series of tapering sections. Diameter outside bark, at both ends of each section, and section length were measured and recorded. Each section was identified to indicate its relative location in the tree and to identify various tree components and product classes. Measurements were taken from ground level to the tip of the main stem, from the base to the tip of each fork, and from the base to the top of all limbs having a base diameter outside bark of one-half inch or more. Therefore, only very small limbs, twigs, and foliage were excluded from above ground biomass measurement.

Each tree's location, species, site, quality, diameter at breast height (d.b.h.), double bark thickness at breast height, and total height were also recorded. Double-bark thicknesses at various

points on the main stem were also measured and recorded for all felled tree samples. Bark-thickness equations by species were developed using upper-stem bark measurements from felled trees and double bark measurements at d.b.h. for all trees. These equations were then used to estimate diameter inside bark at all measurement points throughout the tree. Section volumes were computed with Grosenbaugh's (1952) cubic-foot volume equation, which provides sensitivity to section form and taper. Total-tree volume and total height were obtained by summing all section volumes and summing section lengths in the main stem. All loblolly pine and white oak sample trees were extracted from the large FIA database to provide data for this study.

Southern Loblolly Population

The southern loblolly data set consists of 14,379 loblolly pine (Pinus taeda L.) trees measured in Alabama (data was provided by Roy C. Beltz, Forest Sciences Lab, Starkville, MS). All standing loblolly pine trees encountered in the regular FIA survey were measured, so that trees in the sample were selected proportional to their basal area. The data set contains d.b.h., total height, height to a 4.0″ diameter top, and volume for each tree. A small number of open-growth trees were removed from the data set since morphological differences were expected to exist between these trees and the rest of the data. The volumes were determined using Grosenbaugh's STX program (Grosenbaugh 1964) that calculates volumes from multiple diameter measurements made along the bole. Bole length of all live and salvageable dead trees with d.b.h. of 5.0″ and larger were determined between the top of the one-foot stump and 4.0" diameter outside bark (d.o.b.) or the point where the central stem is terminated by branches, rot, or other disturbances, before reaching 4.0″ d.o.b. On trees which fork below d.b.h., and above 1 foot, the bole length was measured at the fork. The multiple diameter and height measurements used to calculate volume are stump d.o.b. at 1 foot, d.b.h., saw log middle bole d.o.b, sawlog top d.o.b., middle pole d.o.b., and pole top d.o.b. at 4.0″ d.o.b. for saw log sized trees. For pole sized trees, diameters and heights were taken at the stump, d.b.h., middle pole, and pole top to a 4.0″ d.o.b.

3. METHODS

To deduce the error structure of eq. [3] that is most supported by the data, we adapted the "combined variable model" as it is known in forestry, namely $x_i = D_i^2 H_i$ where D_i is the d.b.h. and H_i is the stem height of the i^{th} tree. The widespread use of this model in forestry is due both to its parsimony and to the cogent geometric interpretation it permits.

After exploratory graphical analysis, the following four variance functions, $v_i = f(x_i)$, were advanced as reasonable candidates for further study:

$$\sigma^2 v_i = \sigma^2 x_i^{k_1}, \qquad [4]$$

which has been studied by Cunia (1964), McClure et al.(1983), Meng and Tsai (1986), and Gregoire and Dyer (1989);

$$\sigma^2 v_i = \sigma^2 (x_i + k_2 x_i^{k_3}) \qquad [5]$$

as suggested by Scott, Brewer and Ho (1978);

$$\sigma^2 v_i = \sigma^2 (1 + k_4 x_i + k_5 x_i^{k_6}); \qquad [6]$$

and

$$\sigma^2 v_i = \sigma^2 (1 + k_7 x_i)^2. \qquad [7]$$

The last variance function is appealing since it will be non-negative for any value of the k_7 coefficient. We speculate that the variance of tree volume likewise increase with increasing x_i, but this phenomenon was not pursued since even larger data sets than the one currently available would be needed.

Under model [3] the log-likelihood is

$$\ln L = -\frac{n}{2}\ln(2\pi) + \frac{1}{2}\ln|\Omega^{-1}| - \frac{1}{2}\underline{\epsilon}^T\Omega^{-1}\underline{\epsilon}, \qquad [8]$$

where $\underline{\epsilon}$ is a column vector of residuals $V_i - \alpha - \beta x_i$ and $\Omega = E[\underline{\epsilon}\underline{\epsilon}^T] = \sigma^2$ diag (v_i) where the $v_i's$ are calculated using models [4]-[7]. Ordinary least squares (OLS) estimates of α, β, were used as starting values. Approximate maximum likelihood estimates (MLE's) of α and β under each of the variance models [4]-[7] were obtained using iterative procedures.

The MLE's of the parameter vectors $\underline{\theta}_1 = (\alpha, \beta, \sigma^2, k_1)$, $\underline{\theta}_2 = (\alpha, \beta, \sigma^2, k_2, k_3)$, $\underline{\theta}_3 = (\alpha, \beta, \sigma^2, k_4, k_5, k_6)$, and $\underline{\theta}_4 = (\alpha, \beta, \sigma^2, k_7)$ are those that minimize $-\ln L$ for models [4]-[7], respectively. The minimum of $-\ln L$ with respect to $\underline{\theta}_i$, $i = 1, ...4$ was found using IMSL routine DBCOAH (1989). This constrained optimization routine uses a modified Newton's method with an active set strategy to find $\min -\ln L$ with respect to θ_i $i = 1, ...4$, subject to the constraint $\sigma^2 > 0$. The optimization routine requires both gradient $(\nabla - \ln L)$ and Hessian $(\nabla^2 - \ln L)$ information that must be supplied by the user.

A consistent estimator of the asymptotic variance-covariance matrix is given by the inverse of the information matrix (Kmenta 1986), where the information matrix is the expectation of the Hessian matrix:

$$E[I(\underline{\theta}_i)] = (E[\frac{\delta^2(-\ln L)}{\delta\underline{\theta}_i^2}]) \qquad [9]$$

$i = 1, ...4$. Using the information matrix, confidence intervals of all parameters can be obtained.

Objectives and Criteria for Evaluation

To determine which variance functions best describe the actual variance, an index derived by Furnival (1961) and Cox (1961) is used. This index compares the fit of different variance models and can be written as

$$I = [\text{antilog}\ \frac{\sum_{i=1}^{n}\log_{10}\sqrt{v_i}}{n}]^{-1}S, \qquad [10]$$

where S is the standard error about the model. Smaller values of the index indicate a better fit.

In addition to this index, two different simulation studies were performed to test the performance of the models in estimation.

The first simulation study is entirely model-based. The purpose of the this simulation is to rate the reliability of models[4]-[7] in variance estimation assuming the models. Under the estimated volume model

$$\hat{V}_i = \alpha^* + \beta^* D_i^2 H_i, \qquad [11]$$

the variance of the predicted volume is

$$\text{var}(\hat{V}_i) = \text{var}(\alpha^*) + (D_i^2 H_i)^2\text{var}(\beta^*) + 2D_i^2 H_i\text{cov}(\alpha^*, \beta^*) + \sigma^2, \qquad [12]$$

where $\text{var}(\alpha^*)$, $\text{var}(\beta^*)$, and $\text{cov}(\alpha^*, \beta^*)$, α^*, and β^* were calculated by weighted least squares techniques using models [4]-[7] as weighting functions. Given a simple random sample of size n, the estimate of total population volume is

$$\hat{V} = \frac{N}{n}\sum_{i=1}^{n}\hat{V}_i, \qquad [13]$$

where N is the total number of trees in the population. Thus the variance for the population estimator $\hat{V}$ for a simple random sample is given by

$$\text{var}(\hat{V}) = (\frac{N}{n})^2\sum_{i=1}^{n}\text{var}(\hat{V}_i). \qquad [14]$$

The purpose for the second design-based simulation study was to determine which model provided the best estimate for total volume using model [11]. Simple random sampling was used with estimator [11] to determine if substantial improvements in volume estimation are possible for models [4]-[7], α^* and β^* in [11] were calculated by weighted least squares using models [4]-[7] as weighting functions. The iterated standard errors were calculated for the estimates generated by [13]. In addition, the bias for the estimator given in equation [11] using models [4]-[7] was calculated.

For both simulation studies 10,000 simulations were performed for each of the three populations with models [4]-[7]. The sample size used was $n = 40$. For the first simulation study $\sqrt{\overline{\text{var}(\hat{Y})}}$ was used for comparing models[4]-[7], where $\overline{\text{var}(\hat{Y})}$ was the average variance over the 10,000 simulations. For the second simulation study the iterated standard deviation and bias over the 10,000 simulations was used for comparison.

4. RESULTS AND CONCLUSIONS

The results listed in Table 1 give the coefficients for each variance function, the 95% confidence interval bounds, and the index of fit. For OLS values the index of fit is given. The index of fit indicates that for the southern loblolly data set, the best fitting model was model [5] ($v_i = x_i + k_2 x_i^{k_3}$). For the white oak and southeastern loblolly data set, model [6] ($v_i = 1 + k_4 x_i + k_5 x_i^{k_6}$) was the best fitting model. For both loblolly data sets, model [7], $v_i = (1 + k_7 x_i)^2$, was the worst fitting model. Model [7] was also the second poorest model for the white oak data set. The traditionally used model ($v_i = x_i^{k_1}$) proved to be worst fitting model for the white oak data set and the second poorest fitting model for the loblolly data sets. Comparison of the index for OLS and the four other models show that the latter four differ little but have smaller index values than the OLS solution.

Table 2 contains the results of the two simulation studies for the three populations.

For the model-based simulation there were only small differences between the standard errors for models [4]-[7]. The only conclusion that can be drawn from the model-based simulation study is that all the models seem to be about equally efficient at estimating variance.

For the design-based simulation there was no significant difference between the standard error or bias for models [4]-[7]. The only conclusions that can be drawn from the simulation study are that all the models seem to be about equally efficient at estimating volume using estimator [11]. Note that all four models had substantially smaller standard errors than the OLS solution.

There are a number of practical problems with models [5]-[7] that make them undesirable. Numerically, models [5]-[7] are more difficult to find solutions for the parameter θ_i, $i = 2, ...4$. For the southern loblolly data set, no exact solution could be found for model [6] due to numerical problems with the IMSL routine. The IMSL routine returns an error message stating that the reduction for the $\ln L$ is less than the relative function tolerance. An additional problem with model [6] is that this models can potentially give negative estimates of variance when the linear coefficient k_4 is sufficiently less than zero. The most important problem with models [5]-[7] is that the linear and nonlinear coefficients k_2, k_4, k_5, and k_7 vary drastically across different data sets. This makes it infeasible to recommend a general model that contains more than a single term. Thus model [4] is the most robust of the models studied here.

TABLE 1. Estimates of variance function parameters, 95% confidence intervals, and the index of fit for models (4)-(7).

Model	Southeastern Loblolly	Southern Loblolly	White Oak
(OLS)			
Index	3.7432	5.3417	6.3352
(4) $\sigma^2(x_i^{k_1})$			
σ^2	$0.0000004900 \pm 26.00\%$	$0.0000000641 \pm 14.40\%$	$0.0000001593 \pm .21.50\%$
k_1	$1.8025428088 \pm 1.79\%$	$2.0719429304 \pm 0.84\%$	$1.9764168660 \pm 2.88\%$
Index	1.6158	1.4691	2.0395
(5) $\sigma^2(x_i + k_2 x_i^{k_3})$			
σ^2	$0.0000092859 \pm 76.05\%$	$0.0000355871 \pm 15.40\%$	$0.0000071070 \pm 94.15\%$
k_2	0.0335580189 ± 105.90	$0.0002587115 \pm 46.60\%$	$0.0087936575 \pm 136.13\%$
k_3	$1.8508165307 \pm 2.48\%$	$2.2768744065 \pm 1.60\%$	$2.0786768130 \pm 3.60\%$
Index	1.6148	1.4590	2.0246
(6) $\sigma^2(1 + k_4 x_i + k_5 x_i^{k_6})$			
σ^2	$0.0000245865 \pm 546.50\%$	Failed to Converge	$0.0000709019 \pm 323.98\%$
k_4	$0.0001038379 \pm 3435.45\%$		$0.0340003667 \pm 722.62\%$
k_5	$0.0188104554 \pm 249.33\%$		$0.0010203108 \pm 303.92\%$
k_6	$1.8090922646 \pm 2.60\%$		$2.0635489560 \pm 3.75\%$
Index	1.6156		2.0244
(7) $\sigma^2(1 + k_7 x_i)^2$			
σ^2	$0.0032585491 \pm 36.83\%$	$0.0000098771 \pm 365.77\%$	$0.0000264230 \pm 171.57\%$
k_7	$0.0048798543 \pm 19.58\%$	$0.1085907907 \pm 183.74\%$	$0.0690150212 \pm 86.50\%$
Index	1.6229	1.4725	2.0277

TABLE 2. Standard errors for model-based (MB) and standard errors and bias for design-based (DB) simulations. Standard errors are expressed as a percentage of the model-based and design-based (OLS) standard error. The bias is expressed as a percentage of total tree volume. 10,000 simulations performed.

Model	Southeastern Loblolly	Southern Loblolly	White Oak
(4) $\sigma^2(x_i^{k_1})$			
MB Standard Error	93.8	86.1	78.4
DB Standard Error	79.4	67.2	62.1
DB Bias	-0.76	1.81	0.17
[5] $\sigma^2(x_i + k_2 x_i^{k_3})$			
MB Standard Error	94.4	86.5	79.0
DB Standard Error	79.2	66.8	61.6
DB Bias	-0.79	1.84	0.17
[6] $\sigma^2(1 + k_4 x_i + k_5 x_i^{k_6})$			
MB Standard Error	93.9	No convergence	78.9
DB Standard Error	79.3		61.7
DB Bias	-0.77		0.17
[7] $\sigma^2(1 + k_7 x_i)^2$			
MB Standard Error	96.9	84.9	78.5
DB Standard Error	78.9	67.1	61.9
DB Bias	-0.85	1.59	0.05

4. RECOMMENDATIONS

Since there is little difference in the goodness of fit between the four variance models studied, we recommend variance model (4) where $k_1 \approx 1.8 - 2.1$. For simplicity, the value $k_1 = 2$ is an attractive choice as also noted by Cunia (1964).

REFERENCES

Cost, N. D. 1978. Multiresource inventories: a technique for measuring volumes in standing trees. SW Forest Exp. Sta. USDA For. Serv. Res. Pap. SE-196. 18 p.

Cox, D. R. 1961. Tests of separate families of hypotheses. 4^{th} Berkeley Symp. Proc. 1:101-123.

Cunia, T. 1964. Weighted least squares method and construction of volume tables. For. Sci. 10:180-191.

Davidian, M. and R. J. Carroll. 1987. Variance function estimation. J. Amer. Stat. Assoc. 82,400, 1079-1091.

Furnival, G. M. 1961. An index for comparing equations used in constructing volume tables. For. Sci. 7:337-341.

Gregoire, T. G. and M. E. Dyer. 1989. Model fitting under patterned heterogeneity of variance. For. Sci. 35:105-125.

Grosenbaugh, L. R. 1952. Shortcuts for cruisers and scalers. Southern For. Exp. Sta. USDA For. Serv. Occas. Pap. 126. 24 p.

Grosenbaugh, L. R. 1964. STX-FORTRAN 4 program-for estimates of tree populations from 3P sample-tree-measurements. Pacific SW Forest and Range Exp. Sta. USDA For. Serv. Res. Pap. PSW-13. 49 p.

IMSL. 1989. Math/library. Fortran Subroutines for Mathmematical applications. Version 1.1. IMSL Corp., Houston, Texas.
Kmenta, J. 1986. Elements of econometrics. Ed. 2. MacMillan, New York. 786 p.

McClure, J. P. 1968. Sectional aluminum poles improve length measurements in standing trees. USDA For. Serv. SE For. Exp. Sta. Res. Note SE-98. 4p.

McClure, J. P. 1969. The mirror caliper, a new optical dendrometer. USDA For. Serv. SE For. Exp. Sta. Res. Paper SE-112. 5 p.

McClure, J. P., H. T. Schreuder and R. L. Wilson. 1983. A comparison of several volume table equations for loblolly pine and white oak. USDA For. Serv. SE For. Exp. Sta. Res. Paper SE 240. 8 p.

Meng C. H., W. Y. Tsai. 1986. The selection of weights for a weighted regression of tree volume. Can. J. For. Res. **16**:671- 673.

Scott, A. J., K. R. W. Brewer and E. W. H. Ho. 1978. Finite population sampling and robust estimation. J. Amer. Stat. Assoc. 73,362, 359-361