Corrections for Bias in Regression Estimates After Logarithmic Transformation
Author(s): John J. Beauchamp and Jerry S. Olson
Source: *Ecology*, Vol. 54, No. 6 (Nov., 1973), pp. 1403-1407
Published by: Ecological Society of America
Stable URL: http://www.jstor.org/stable/1934208
Accessed: 09/12/2009 19:11

CORRECTIONS FOR BIAS IN REGRESSION ESTIMATES AFTER LOGARITHMIC TRANSFORMATION[1]

JOHN J. BEAUCHAMP
Mathematics Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37830

AND

JERRY S. OLSON
Environmental Sciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37830

Abstract. Experience with biological data, such as dimensions of organisms, often confirms that logarithmic transformations should precede the testing of hypotheses about regression relations. However, estimates also may be needed in terms of untransformed variables. Just taking antilogarithms of values from a log-log regression line or function leads to biased estimates. This note compares corrections for this bias, and includes an example relating mass of tree parts (bole, branches, and leaves) to tree diameter of tulip poplar (*Liriodendron tulipifera* L.) in Oak Ridge, Tennessee, forests. An Appendix summarizes derivation of exact and approximate unbiased estimators of expected values from log-antilog regression, and of variance around the unbiased regression line.

INTRODUCTION

Regression analysis of the dimensions of organisms can readily be accomplished by computer programs using either the original measurements of organisms or various transformations of the raw data. In many cases the variability around a fitted line increases in proportion to the mean size. The variability may be stabilized by taking the log transformation of the data and the transformed data more closely satisfy the assumptions underlying most parametric statistical methods, such as regression and analysis of variance. Tests of hypotheses, which may be used in deciding about the desirability of pooling parts of the total data, are then valid. Baskerville (1970) provides a diagnostic program to help guide the user's judgement leading up to his decisions about such analysis, and about subsequent use of the resulting data.

In cases like the example below, there is also a need for estimates in terms of the original scale, so that results can be combined to obtain an estimate of a quantity like total mass per unit area of forest. Estimates, which are obtained by taking antilogs of the previously transformed data, are common (Ovington and Olson 1970). Nevertheless, a bias is inherent in this procedure because the largest values are compressed on the logarithmic scale and thereby tend to have less "leverage" than small values in making such an estimate.

Successive approximations to correct for this bias were long ago outlined by Finney (1941), but are seldom used in practice (Madgwick 1970). Aitchison and Brown (1969) discuss many of the estimation problems and properties associated with a random variable Z whose logarithm is normally distributed. Finney's approach has been adapted to log-normal regression by Mostafa and Mahmoud (1964), who carry the series approximation to terms of order $1/n$ where n is the sample size. The present note carries similar approximations to terms of $1/n^2$ and also extends Laurent's (1963) approach which used modified Bessel functions for minimum variance unbiased estimates of the median and variance. Heinen (1968) and Bradu and Mundlak (1970) use infinite series for unbiased estimates of the mean in lognormal regression. Zellner (1971) uses estimators for log-linear regression which are optimal in the Bayesian sense, and also reviews some non-Bayesian results. The following sections summarize point estimation for the mean and variance around linear regression, and may be useful with or without confidence interval estimates like those of Land (1972). Such a correction for the exact minimum variance unbiased estimate given in the Appendix of this note is likely to prove important in forestry, several aspects of production ecology and the allometric analysis of growth and form.

ESTIMATION PROCEDURE

Symbolically, the problem is the consideration of the random variable Z such that $Y = \ln Z$ is normally distributed with mean, $E(Y) = \beta_0 + \beta_1 x$, which is a linear function of an independent-nonrandom vari-

[1] Contribution No. 83 from the Eastern Deciduous Forest Biome, US-IBP. Research supported in part by the U. S. Atomic Energy Commission under contract with the Union Carbide Corporation, and in part by the Deciduous Forest Biome Project, International Biological Program (IBP), funded by the National Science Foundation under Interagency Agreement AG-199, 40-193-69 with the Atomic Energy Commission–Oak Ridge National Laboratory. Manuscript received October 24, 1972; accepted May 11, 1973.

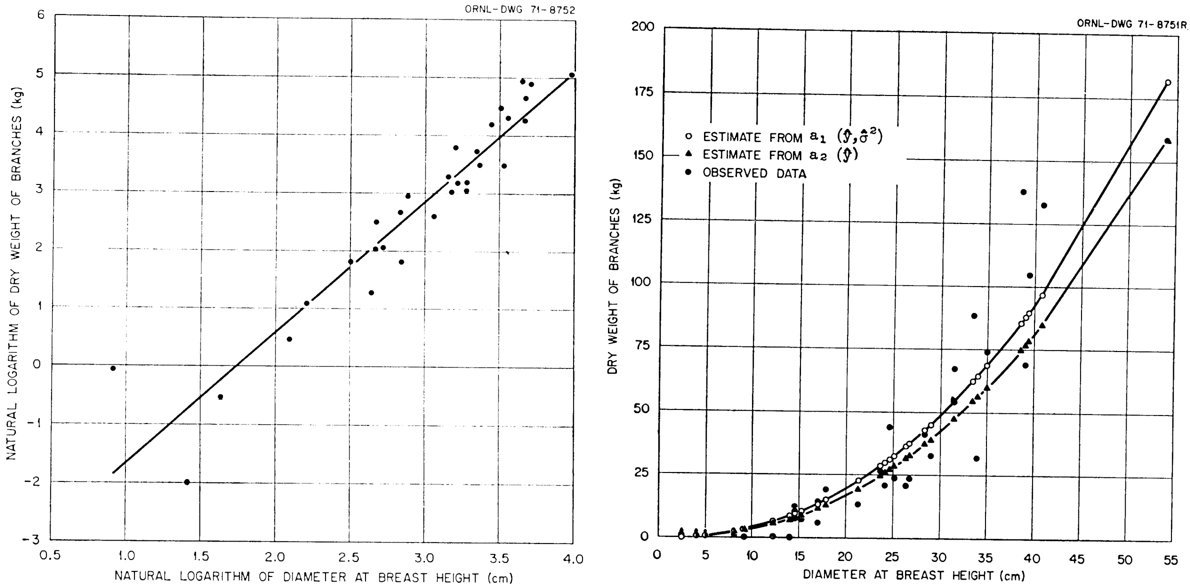

FIG. 1. Comparisons of log-log regression for *Liriodendron tulipifera* L. branch weight on bole diameter (a, left) with observed data, an unbiased estimate a_1 and a biased estimate a_2 obtained by using antilogarithms without correction (b, right).

able x, and with variance σ^2. From a sample of n observations on Z, denoted by $z_1, z_2, \ldots, z_n$, it is desired to obtain minimum variance unbiased estimates of the mean and variance of the distribution of Z. A derivation of the estimation procedure is given in the Appendix and the steps involved are outlined below:

Step 1: For each value of z_i there is a corresponding pair of values (x_i, y_i) for $i = 1, 2, \ldots, n$. Fit a straight line to the n pairs (x_i, y_i) to obtain the least-squares estimates of β_0 and β_1, denoted by $\hat{\beta}_0$ and $\hat{\beta}_1$, respectively.

Step 2: From the regression analysis of the n pairs (x_i, y_i), calculate $\hat{\sigma}^2$ from the expression given in the Appendix.

Step 3: For each value of the independent variable calculate ϕ as given in the Appendix.

Step 4: From the above calculations the minimum variance unbiased estimator of Z is given by $a_1(\hat{y}, \hat{\sigma}^2) = \exp(\hat{\beta}_0 + \hat{\beta}_1 x)\psi(\hat{\sigma}^2/2)$ where $\psi(\hat{\sigma}^2/2)$ is defined in equation (1) of the Appendix.

It should be noted that the first term, $\exp(\hat{\beta}_0 + \hat{\beta}_1 x)$, in $a_1(\hat{y}, \hat{\sigma}^2)$ is the estimate one would obtain by taking the antilog of $\hat{y}$, which is a biased estimate of the mean of Z. Therefore, $\psi(\hat{\sigma}^2/2)$ is the term which corrects for the bias when only $\exp(\hat{\beta}_0 + \hat{\beta}_1 x)$ is used. If expressions for the modified Bessel function in $\psi(\hat{\sigma}^2/2)$ are not available, then equation (3) of the Appendix may be used. This series expression is an approximation to the unbiased estimator of the mean of Z through terms of order $1/n^2$ and has been a very close approximation in practical examples like that given in the next section.

Step 5: Approximations to the unbiased estimator of the variance of Z, through terms of order $1/n^2$, and variance of $a_1(\hat{y}, \hat{\sigma}^2)$, to terms of order $1/n$ may be found by using equations (5) and (8) in the Appendix.

The computer program to make these calculations includes the successive approximations of numerical data and several options of graphical output (Beauchamp, Hull, and Olson 1972).

EXAMPLE

In this section the estimation procedure of the previous section is applied to the data on tulip poplar, *Liriodendron tulipifera*, given in Table 1. For this particular example x is the natural logarithm of the diameter (cm) of the tree at breast height, and Y is the natural logarithm of the dry weight (kg) of the branches. Although there are errors present in both measurements, it is assumed that the error in the measurement of the diameter is negligible in comparison to the measurement error plus random variations of the dry weight of all the branches in a population of trees having any given diameter.

The estimates of β_0, β_1, and σ^2, needed in Steps 1 and 2 mentioned above, were obtained from the regression analysis of Y on x and are given by $\hat{\beta}_0 = -3.9195$, $\hat{\beta}_1 = 2.2536$, and $\hat{\sigma}^2 = 0.2971$. Figure 1(a) shows the observed transformed data and the line

TABLE 1. Data and results from lognormal regression for branch weight of *Liriodendron tulipifera* L.

Tree diameter[a] (cm)	Dry wt. branches[b] (kg)	Unadjusted estimated mean[c]	$\psi(\hat{\sigma}^2/2)$[d]	Adjusted mean[e]	Estimate of variance[f]
(1)	(2)	(3)	(4)	(5)	(6)
2.5	.94238	.15652	1.1083	.17346	.00304
4.1	.13423	.47724	1.1274	.53806	.01770
5.1	.58165	.78046	1.1343	.88526	.03731
8.1	1.5963	2.2138	1.1455	2.5358	.16969
9.1	2.9754	2.8779	1.1476	3.3025	.24584
12.2	6.1073	5.5718	1.1515	6.4159	.63253
14	3.557	7.5978	1.1527	8.7579	1.0133
14.5	7.680	8.2231	1.1529	9.4806	1.1490
14.5	11.982	8.2231	1.1529	9.4806	1.1490
15.2	7.718	9.1449	1.1532	10.5459	1.3664
17	14.300	11.7683	1.1537	13.5767	2.1127
17	6.0178	11.7683	1.1537	13.5767	2.1127
17.8	18.905	13.0534	1.1538	15.0606	2.5556
21.3	13.333	19.5620	1.1537	22.5696	5.7618
23.6	26.598	24.6474	1.1534	28.4286	9.6259
24.1	20.95	25.8399	1.1533	29.8015	10.733
24.6	44.138	27.0638	1.1532	31.2102	11.954
25.1	23.725	28.3193	1.1531	32.6549	13.299
26.4	20.71	31.7325	1.1528	36.5806	17.456
26.7	23.61	32.5509	1.1527	37.5215	18.566
28.4	41.13	37.4090	1.1522	43.1033	26.113
29	32.66	39.2137	1.1520	45.1756	29.355
31.5	67.359	47.2466	1.1512	54.3912	46.913
33.5	88.584	54.2774	1.1505	62.4469	66.821
34	31.73	56.1202	1.1503	64.5568	72.790
35	74.21	59.9087	1.1500	68.8927	86.091
38.6	137.92	74.6985	1.1486	85.7966	152.15
39.1	69.22	76.8968	1.1484	88.3063	164.01
39.4	104.79	78.2328	1.1483	89.8312	171.49
40.8	132.52	84.6374	1.1477	97.1378	210.23
54	159.00	159.185	1.1422	181.825	1.072.4
Sum of squared residuals		10,180.2		8.173.7	

[a] Observed at breast height: 4.5 feet = 137 cm.
[b] Observed, dried at 105° C.
[c] From $a_2(\hat{y})$.
[d] From equation (1).
[e] From $a_1(\hat{y}, \sigma^2)$.
[f] From equation (8).

$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x$. The values of the unadjusted biased mean $a_2(\hat{y}) = \exp(\hat{\beta}_0 + \hat{\beta}_1 x)$, which are the values obtained by merely taking the antilogs of the transformed estimates, are given in column 3 of Table 1. The values of $\psi(\hat{\sigma}^2/2)$, needed to obtain $a_1(\hat{y}, \hat{\sigma}^2)$, are given in column 4 of Table 1. By multipling the values in columns 3 and 4 together, the minimum variance unbiased estimate $a_1(\hat{y}, \hat{\sigma}^2)$, is obtained and its values are given in column 5. The estimated variance of $a_1(\hat{y}, \hat{\sigma}^2)$, mentioned in Step 5, is given in column 6 of Table 1. Figure 1(b) shows the observed data with the fitted curves from the expressions for $a_1(\hat{y}, \hat{\sigma}^2)$ and $a_2(\hat{y})$. A closer examination of data for the two curves in Figure 1(b) shows dry weight estimated from $a_1(\hat{y}, \hat{\sigma}^2)$ is from 10 to 13 percent greater than the dry weight estimated from $a_2(\hat{y})$ for a few small trees (2.5 to ~5 cm diameter). These trees would make very little contribution to the total branch mass per unit area in a forest having many tree sizes. Over a wide range of tree diameters (8 to 54 cm) the difference between these two estimates is about 15 percent.

Data from boles for the same trees showed smaller bias (<1 percent) in the estimated dry weight obtained from $a_2(\hat{y})$ relative to our proposed estimate $a_1(\hat{y}, \hat{\sigma}^2)$. For 16 of the trees mentioned above, leaves were sampled during the growing season; these showed differences between the two estimates near 6%. In many forests, foliage represents a smaller fraction of the total mass than either branches or boles, but can be important in nutrient budgets and exchange. These and other examples not shown here thus suggest there are some data for which our refinements and those suggested by authors cited in the first paragraphs are negligible for practical purposes. However, for other cases, the additional terms in expression (3) of the Appendix of the estimate $a_1(\hat{y}, \hat{\sigma}^2)$ could compensate for the distortion introduced by transforming raw data and retransforming logarithmic regression estimates into terms of the original observations.

DISCUSSION

In this paper the authors have used the properties of the lognormal distribution to obtain unbiased estimates for the mean and variance of lognormal variates in the regression framework which are improvements over simply taking antilogarithms after the logarithmic transformation. Bakersville (1973) has investigated estimates obtained by substituting $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\sigma}^2$ in $E(Z_i) = \exp\{\beta_0 + \beta_1 x_i + \sigma^2/2\}$ for approximating the mean of the lognormally distributed random variable at each particular value of the independent variable. The above estimates of the mean are still biased since the true β_0, β_1, and σ^2 are unknown. Therefore the development of the estimates in Section 2 and the Appendix is a way to estimate and eliminate the bias. Further discussion of the foregoing example and others indicates that Baskerville's approximation already may be close to the unbiased value unless the variance is quite large.

By simply dropping all terms except "1" in the parentheses of the approximation in (3) of the Appendix, one obtains the estimator proposed by Baskerville (1973). The additional parenthetical terms in (3) would make contributions depending on the magnitude of $\hat{\sigma}^2$ and ϕ, and also give an indication of the bias in the estimator $\exp\{\hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\sigma}^2/2\}$. The computer program mentioned earlier (Beauchamp, Hull, and Olson 1972) includes in the output a comparison of the following estimators: (1) Y EST $= \exp\{\hat{\beta}_0 + \hat{\beta}_1 x_i\}$; (2) $Y0$ EST $= \exp\{\hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\sigma}^2/2\}$; (3) series ($Y1$ EST; $Y2$ EST) approximations to the unbiased estimator through terms of order $1/n$ and $1/n^2$; and (4) the unbiased estimator $Y3$ EST $= a_1(\hat{y}, \hat{\sigma}^2)$. For any particular set of data it may be helpful to compare the estimates which one obtains from these different methods in order to determine the improvement of one estimator over another.

Figure 2 shows estimates from the data of Table 1 replotted so the solid diagonal line gives the unbiased

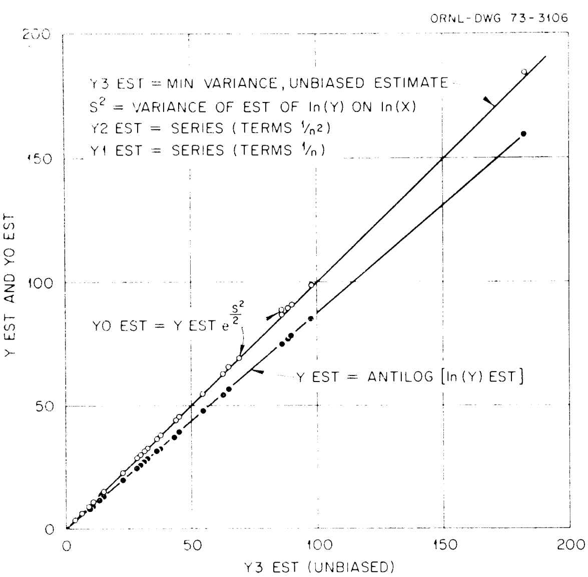

FIG. 2. Comparison of estimators in lognormal regression.

estimate. As in Figure 1, Y EST (sample dots and interpolated dashes) shows a bias which could be of serious practical concern.

However, $Y0$ EST is already a much closer approximation to $Y3$ EST. $Y1$ EST and $Y2$ EST are not even plotted because they could not be distinguished from the diagonal line on the present scale of plotting. A related way of expressing the improvement in estimates is that the sum of squares of residual variation around the several estimated fits in our example diminished from 10,180 for Y EST (antilog unadjusted) to 8,180 for $Y0$ EST, to 8,174 for $Y1$ EST, $Y2$ EST and $Y3$ EST.

ACKNOWLEDGMENTS

The authors are indebted to several referees for suggestions in review of the manuscript. The FORTRAN IV program is available on request from the authors.

LITERATURE CITED

Aitchison, J., and J. A. C. Brown. 1969. *The lognormal distribution.* Cambridge University Press, Cambridge, U. K.

Baskerville, G. L. 1970. Testing the uniformity of variance in arithmetic and logarithmic units of a Y-variable for classes of an X-variable. USAEC Report ORNL-IBP-70-1, Oak Ridge National Laboratory, Oak Ridge, Tn.

———. 1973. Use of the logarithmic equation in the estimation of plant biomass. Can. J. Forestry (in press).

Beauchamp, J. J., Norma C. Hull, and J. S. Olson. 1972. LØGNØRM: Computer program for unbiased allometric regression estimates. Eastern Deciduous Forest Biome Memo Report No. 71-102, Oak Ridge National Laboratory, Oak Ridge, Tn.

Beauchamp, J. J., and J. S. Olson. 1972. Estimators for the mean and variance of a lognormal distribution where the mean is a function of an independent variable. Eastern Deciduous Forest Biome Memo Report No. 71-101, Oak Ridge, Tn.

Bradu, D., and Y. Mundlak. 1970. Estimation in lognormal linear models. J. Amer. Statist. Assoc. **65:** 198–211.

Finney, D. J. 1941. On the distribution of a variate whose logarithm is normally distributed. J. R. Statist. Soc. Supplement. **7:** 155–161.

Gradshteyn, I. S., and I. M. Ryzhik. 1965. Table of Integrals, Series, and Products. Academic Press, New York. 1095.

Heien, D. M. 1968. A note on log-linear regression. J. Amer. Statist. Assoc. **63:** 1034–1038.

Land, C. E. 1972. An evaluation of approximate confidence interval estimation methods for lognormal means. Technometrics **14:** 145–158.

Laurent, A. G. 1963. The lognormal distribution and the translation method: description and estimation problems. J. Amer. Statist. Assoc. **58:** 231–235.

Madgwick, H. A. I. 1970. Biomass and productivity models of forest canopies. *In* D. E. Reichle [ed.] Ecological Studies 1. Analysis of Temperate Forest Ecosystems. pp. 47–54. Springer-Verlag, New York, Heidelberg and Berlin.

Mostafa, M. D., and M. W. Mahmoud. 1964. On the problem of estimation for the bivariate lognormal distribution. Biometrika **51:** 522–527.

Ovington, J. D., and J. S. Olson. 1970. Biomass and chemical content of El Verde Lower Montane Rain Forest plants. *In* H. T. Odum and R. F. Pigeon [eds.] A tropical rainforest. H-53-77. U. S. Atomic Energy Commission, Div Technical Information. TID-24270. (Clearinghouse Fed. Sci. Tech. Info., Springfield, Va. 22151.)

Zellner, A. 1971. Bayesian and non-Bayesian analysis of the lognormal distribution and lognormal regression. J. Amer. Statist. Assoc. **66:** 327–330.

APPENDIX

Derivation of estimators

Consider the random variable Z such that $Y = \ln Z$ is normally distributed with mean, $E(Y) = \beta_0 + \beta_1 x$, which is a linear function of an independent-nonrandom variable x, and with variance σ^2. From a sample of n observations on Z, desire minimum variance unbiased estimates of the mean and variance of the distribution of Z. If z_i represents the corresponding observed value of Z for $i = 1, 2, \ldots, n$, then by using properties of the lognormal distribution $E(z_i) = \exp\{\beta_0 + \beta_1 x_i + \sigma^2/2\}$ and $\mathrm{Var}(z_i) = (\exp(\sigma^2) - 1) \exp(2\beta_0 + 2\beta_1 x_i + \sigma^2)$.

Let $\hat{\beta}_0$ and $\hat{\beta}_1$ be the maximum likelihood estimators of the paramenters β_0 and β_1, respectively, obtained from the n (x_i, y_i) pairs of observations transformed so that $y_i = \ln z_i$. Then the predicted value $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ is normally distributed with mean $\beta_0 + \beta_1 x$ and variance $\sigma^2\phi/n$; where $\phi = \sum_{i=1}^{n} (x_i - x)^2 / \sum_{i=1}^{n} (x_1 - \bar{x})^2$.

By following an approach similar to that provided by Finney (1941), a minimum variance unbiased estimator of $E(Z)$ is given by $a_1(\hat{y}, \hat{\sigma}^2) = \exp(\hat{\beta}_0 + \hat{\beta}_1 x)\psi(\hat{\sigma}^2/2)$. In this expression $\hat{\sigma}^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2/(n-2)$ is an unbiased estimator of σ^2 which is independent of $\hat{\beta}_0$ and $\hat{\beta}_1$;

$$\psi(t) = \frac{\Gamma((n-2)/2)}{[(1-\phi/n)(n-2)t/2]^{(n-4)/4}} \times$$

$$I_{(n-4)/2}(\sqrt{2(1-\phi/n)(n-2)t}); \tag{1}$$

and the generalized expression of the last term in (1)

$$I_\nu(u) = \sum_{r=0}^{\infty} [r!\Gamma(\nu+r+1)]^{-1}(u/2)^{\nu+2r} \tag{2}$$

is a modified Bessel function. If evaluations of the modified Bessel function are not available, by expanding $\psi(\hat{\sigma}^2/2)$ in increasing powers of $1/n$, a good approximation to the unbiased estimator of $E(Z)$ through terms of order $1/n^2$ is given by

$$\exp(\hat{\beta}_0 + \hat{\beta}_1 x + \hat{\sigma}^2/2)\left\{1 - \frac{\hat{\sigma}^2(2\phi + \hat{\sigma}^2)}{4n} + \frac{(\hat{\sigma}^2)^2[(\hat{\sigma}^2)^2 + 2(10/3 + 2\phi)\hat{\sigma}^2 + 4\phi^2 + 16\phi]}{32n^2}\right\}. \tag{3}$$

From the expressions for $E[\psi(r^2\hat{\sigma}^2/2)]$ and $E[\exp(2\hat{\beta}_0 + 2\hat{\beta}_1 x)]$, an unbiased estimator of $\mathrm{Var}(Z)$ is given by

$$\left\{\psi(2\hat{\sigma}^2) - \psi\left(\frac{1-2\phi/n}{1-\phi/n}\hat{\sigma}^2\right)\right\}\exp(2\hat{\beta}_0 + 2\hat{\beta}_1 x). \tag{4}$$

By expanding $\psi(t)$ an approximation to the unbiased estimator of $\mathrm{Var}(Z)$ through terms of order $1/n^2$ is then given by

$$\{\exp(2\hat{\beta}_0 + 2\hat{\beta}_1 x + \hat{\sigma}^2)\}\left\{(\exp\hat{\sigma}^2)\left[1 - \frac{2\hat{\sigma}^2(\phi + 2\hat{\sigma}^2)}{n} + \frac{2(\hat{\sigma}^2)^2[4(\hat{\sigma}^2)^2 + 2(10/3 + 2\phi)\hat{\sigma}^2 + \phi^2 + 4\phi]}{n^2}\right]\right.$$
$$\left. - \left(1 - \frac{\hat{\sigma}^2}{n}(\hat{\sigma}^2 + 2\phi) + \frac{(\hat{\sigma}^2)^2}{2n^2} \times [(\hat{\sigma}^2)^2 + (10/3 + 4\phi)\hat{\sigma}^2 + 4\phi^2 + 8\phi]\right)\right\}. \tag{5}$$

Since the minimum variance unbiased estimator of $E(Z)$ involves the product of two functions which are independent, after extensive algebra (Beauchamp and Olson 1972) we find

$$\mathrm{Var}[a_1(\hat{y}, \hat{\sigma}^2)] = e^{2\beta_0 + 2\beta_1 x + 2\sigma^2\phi/n}\left\{\frac{2^{3/2}\Gamma(n-2)\Gamma((n-3)/2)}{\Gamma((n-1)/2)\Gamma(n-3)}\,{}_2F_2\left[\frac{n-3}{2}, \frac{n-2}{2}; \frac{n-2}{2}, n-3; 2(1-\phi/n)\sigma^2\right] - e^{(1-\phi/n)\sigma^2}\right\}$$
$$+ e^{2\beta_0 + 2\beta_1 x + (1-\phi/n)\sigma^2}(e^{2\sigma^2\phi/n} - e^{\sigma^2\phi/n}). \tag{6}$$

Here

$${}_pF_q(\alpha_1, \ldots, \alpha_p; \nu_1, \ldots, \nu_q; u) = \sum_{R=0}^{\infty} \frac{(\alpha_1)_R(\alpha_2)_R \ldots (\alpha_p)_R}{(\nu_1)_R(\nu_2)_R \ldots (\nu_q)_R}\frac{u^R}{R!} \tag{7}$$

is a generalized hypergeometric series as defined in Gradshteyn and Ryzhik (1965) with $(\alpha_i)_R = \Gamma(\alpha_i + R)/\Gamma(\alpha_i)$. Since it may not be easy to evaluate the generalized hypergeometric series, the following expression is given as an approximation to $\mathrm{Var}[a_1(\hat{y}, \sigma^2)]$ to terms of order $1/n$;

$$\left(\frac{\sigma^2\phi}{n} + \frac{\sigma^4}{2n}\right)\exp[2(\beta_0 + \beta_1 x) + \hat{\sigma}^2]. \tag{8}$$