Smearing Estimate: A Nonparametric Retransformation Method
Author(s): Naihua Duan
Source: *Journal of the American Statistical Association*, Vol. 78, No. 383 (Sep., 1983), pp. 605-610
Published by: Taylor & Francis, Ltd. on behalf of the American Statistical Association
Stable URL: https://www.jstor.org/stable/2288126
Accessed: 17-01-2026 13:54 UTC

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.

Your use of the JSTOR archive indicates your acceptance of the Terms & Conditions of Use, available at

Taylor & Francis, Ltd., *American Statistical Association* are collaborating with JSTOR to digitize, preserve and extend access to *Journal of the American Statistical Association*

Smearing Estimate: A Nonparametric Retransformation Method

NAIHUA DUAN*

The smearing estimate is proposed as a nonparametric estimate of the expected response on the untransformed scale after fitting a linear regression model on a transformed scale. The estimate is consistent under mild regularity conditions, and usually attains high efficiency relative to parametric estimates. It can be viewed as a low-premium insurance policy against departures from parametric distributional assumptions. A real-world example of predicting medical expenditures shows that the smearing estimate can outperform parametric estimates even when the parametric assumption is nearly satisfied.

KEY WORDS: Retransformation; Transformation; Nonparametric; Prediction; Lognormal linear model; Cobb-Douglas function.

1. INTRODUCTION

A monotonic transformation is often applied to observations recorded on an untransformed scale to achieve desirable statistical properties such as additivity, homoscedasticity, and normality. Certain analyses (e.g., fitting a least squares regression model) are carried out on the transformed scale, possibly combined with certain inferences such as significance tests on comparisons of experimental treatments. However, it is also very often desirable to carry out certain procedures, such as prediction and forecasting, on the untransformed scale. In doing so, one will be confronted with the problem of retransformation bias; namely, unbiased and consistent quantities on the transformed scale usually do not retransform into unbiased or consistent quantities on the untransformed scale.

In this article, we propose a nonparametric method, the *smearing estimate*, as an estimate of an individual's expected response on the untransformed scale. (The terminology "smearing" was originally coined by C. Morris for the tactic of distributing (smearing) the excess in one observation to other observations proportionally when adjusting unlogged median estimates to unlogged mean estimates.) The essence of the procedure is to estimate the unknown error distribution by the empirical cdf of the estimated regression residuals, and then take the desired expectation with respect to the estimated error distribution. In a broader context, the method can be viewed as an application of the bootstrap principle (Efron 1979).

In the next section, we present the retransformation problem, and use an example to demonstrate the possible bias due to inappropriate use of the normal assumption. In the third section, we derive the smearing estimate as an estimate of the untransformed scale expectation free from distributional assumptions on the error distribution F. The consistency property of the smearing estimate is established in Section 4. In Section 5, we examine the efficiency of the smearing estimate compared with a parametric estimate when the parametric assumption is satisfied. In the last section, we discuss an application of the smearing estimate to a real-world prediction problem, namely, the Rand Health Insurance Study (HIS), for which this methodology was derived.

2. THE RETRANSFORMATION PROBLEM

We denote the observations on the untransformed scale by Y_i, $i = 1, \ldots, n$, the transformed observations by η_i, $i = 1, \ldots, n$, which are related by

$$\eta_i = g(Y_i),\ Y_i = h(\eta_i),\ h = g^{-1},$$

where g and h are assumed to be monotonic and continuously differentiable. To avoid the trivial cases, we also assume g and h to be nonlinear. We refer to g as the *transformation* and to h as the *retransformation*. We assume g and h to be known a priori.

We consider a linear regression model on the transformed scale:

$$\eta_i = x_i\beta + \epsilon_i,$$

$$\epsilon_i \sim F \text{ (iid)},\ \mathrm{E}\,\epsilon_i = 0,\ \mathrm{var}\,\epsilon_i = \sigma^2,$$

where x_i's are given row vectors of explanatory variables, β is a column vector of unknown parameters to be estimated, and ϵ_i's are the residual errors. Although the error distribution F is usually assumed to be normal, we do not make this assumption. We show later in this section that inappropriate use of the normal assumption can lead to inconsistent prediction results.

---

* Naihua Duan is Associate Statistician, Economics Department, the Rand Corporation, 1700 Main Street, Santa Monica, CA 90406. This research was supported in part by the Health Insurance Study, Grant 016B80 from the Department of Health and Human Services, and in part by Rand corporate research funds. The views expressed herein do not necessarily reflect those of the Rand Corporation or the Department of Health and Human Services. The author wishes to thank J. Acton, B. Efron, L. Lillard, W. G. Manning, C. N. Morris, J. P. Newhouse, W. R. Ott, W. H. Rogers, J. E. Rolph, J. Smith, and the referees for helpful comments and discussions.

© Journal of the American Statistical Association
September 1983, Volume 78, Number 383
Theory and Methods Section

For the assumed model, the minimum variance linear unbiased estimate of β is the least squares regression estimate on the transformed scale: $\hat{\beta} = (X'X)^{-1}X'\eta$, where $X = (x_1', \ldots, x_n')$ is the design matrix, assumed to have full rank, and $\eta = (\eta_1, \ldots, \eta_n)'$ is the transformed data vector. Moreover, for an individual with explanatory variables x_0, the prediction $x_0\hat{\beta} = x_0(X'X)^{-1}X'\eta$ is the minimum variance linear unbiased estimate of the expectation of his response $\mathrm{E}\,\eta_0 = x_0\beta$ on the transformed scale. Moreover, the regression coefficients $\hat{\beta}$, as well as the prediction $x_0\hat{\beta}$ for fixed x_0, are unbiased and also consistent if the design matrix is asymptotically nondegenerate.

In terms of the untransformed scale, it may seem natural to retransform the transformed scale prediction $x_0\hat{\beta}$ by $h = g^{-1}$, and use $h(x_0\hat{\beta})$ to estimate the expectation of the individual's response $\mathrm{E}\,Y_0 = \mathrm{E}\,h(\eta_0) = \mathrm{E}\,h(x_0\beta + \epsilon)$ on the untransformed scale. However, the prediction $h(x_0\hat{\beta})$ will be neither unbiased nor consistent unless the transformation is linear, which we have assumed is not the case. Actually, even if we know the true parameters β, $h(x_0\beta)$ is not the correct "estimate" of $\mathrm{E}\,Y_0$:

$$\mathrm{E}\,Y_0 = \mathrm{E}\,h(x_0\beta + \epsilon) \neq h(x_0\beta).$$

There is extensive literature (e.g., Neyman and Scott 1960; Meulenberg 1965; Bradu and Mundlak 1970; Ebbeler 1973; Mehran 1973; Evans and Shaban 1974; Shimizu and Iwase 1981) devoted to the problem of estimating the untransformed scale expectation under the assumption that the error distribution is normal. We refer to those results categorically as *normal theory estimates*.

It should be noted that the normality assumption plays a very different role in estimating the untransformed scale expectations than in estimating the regression coefficients. For estimating the regression coefficients, whether the true error distribution is normal or not, the least squares estimate, which is the maximum likelihood estimate under the normal assumption, is consistent and minimum variance linear unbiased. When the true error distribution is not normal, the normality assumption affects only the efficiency of our estimate (Cox and Hinkley 1968). If we know the form of the true error distribution, we can sometimes derive alternative estimates that are more efficient than the least squares estimate. However, for estimating the untransformed scale expectation, an incorrect normality assumption can lead to inconsistent estimates.

For example, in the case of a logarithmic transformation with normally distributed error, the untransformed scale expectation is $\exp(x_0\beta + \sigma^2/2)$, where $\sigma^2 = \operatorname{var}\epsilon$. The expectation can be estimated consistently by any of the normal theory estimates, such as the naive estimate $\exp(x_0\hat{\beta} + \hat{\sigma}^2/2)$, where $\hat{\beta}$ denotes the least squares regression coefficients and $\hat{\sigma}^2$ denotes the mean squared error.

Whether the true error distribution is normal or not, the above estimate is consistent for $\exp(x_0\beta + \sigma^2/2)$; however, it might not be consistent for $\mathrm{E}\,Y_0$. For example, if the true error distribution is actually a mixture of two normal distributions,

$$\begin{aligned} &\epsilon \sim N(0, .95\sigma^2) \quad \text{with probability } .995, \\ &\epsilon \sim N(0, 10.95\sigma^2) \quad \text{with probability } .005, \\ &\quad (\operatorname{var}\epsilon = \sigma^2), \end{aligned}$$

then the untransformed scale expectation is

$$\begin{aligned} \mathrm{E}\,Y_0 = {} & .995 \exp(x_0\beta + .475\,\sigma^2) \\ & + .005 \exp(x_0\beta + 5.475\,\sigma^2). \end{aligned}$$

For $\sigma^2 = 1$, we have $\mathrm{E}\,Y_0 = 2.79 \exp(x_0\beta)$. The normal theory estimates converge almost surely to $\exp(x_0\beta + \sigma^2/2) = 1.65 \exp(x_0\beta)$, which has a 41 percent asymptotic bias.

3. THE SMEARING ESTIMATE

Our goal is to estimate the untransformed scale expectation

$$\mathrm{E}\,Y_0 = \mathrm{E}\,h(x_0\beta + \epsilon) = \int h(x_0\beta + \epsilon)\,dF(\epsilon).$$

Without knowing the error distribution function F or a reliable parametric form for it, we estimate F by the empirical cdf of the estimated residuals

$$\hat{F}_n(e) = \frac{1}{n}\sum_{i=1}^{n} I\{\hat{\epsilon}_i \leq e\},$$

where $\hat{\epsilon}_i = \eta_i - x_i\hat{\beta}$ denotes the least squares residual, $I\{\cdot\}$ denotes the indicator function of the event "$\cdot$".

As is usual in nonparametric analyses, a population quantity with an expression in terms of the true cdf can be estimated by the corresponding expression in terms of the empirical cdf. For example, the population mean $\mu = \int x\,dF(x)$ can be estimated nonparametrically by the sample mean $\bar{x} = \int x\,d\hat{F}_n(x)$. Similarly, we estimate $\mathrm{E}\,Y_0$ by substituting the unknown cdf F by its empirical estimate $\hat{F}_n$:

$$\begin{aligned} \tilde{\mathrm{E}}\,Y_0 &= \int h(x_0\beta + \epsilon)\,d\hat{F}_n(\epsilon) \\ &= \frac{1}{n}\sum_{i=1}^{n} h(x_0\beta + \hat{\epsilon}_i). \end{aligned}$$

Further substituting the regression parameters β by their least squares estimates $\hat{\beta}$, we have the estimate

$$\begin{aligned} \hat{\mathrm{E}}\,Y_0 &= \int h(x_0\hat{\beta} + \epsilon)\,d\hat{F}_n(\epsilon) \\ &= \frac{1}{n}\sum_{i=1}^{n} h(x_0\hat{\beta} + \hat{\epsilon}_i), \end{aligned} \tag{3.1}$$

which we refer to as the *smearing estimate*.

4. CONSISTENCY OF THE SMEARING ESTIMATE

Assuming that h is continuously differentiable, we take the first-order Taylor expansion:

$$\begin{aligned} h(x_0\hat{\beta} + \hat{\epsilon}_i) = {} & h(x_0\beta + \epsilon_i) \\ & + \delta_i \times h'\,(x_0\beta + \epsilon_i + \theta_i\delta_i), \end{aligned}$$

where

$$0 \le \theta_i \le 1,$$
$$\delta_i = (x_0\hat{\beta} + \hat{\epsilon}_i) - (x_0\beta + \epsilon_i)$$
$$= (x_0 - x_i)(X'X)^{-1}X'\epsilon.$$

The smearing estimate can be decomposed as follows:

$$\hat{E}\ Y_0 = \frac{1}{n}\sum_{i=1}^{n} h(x_0\hat{\beta} + \hat{\epsilon}_i)$$
$$= \frac{1}{n}\sum_{i=1}^{n} h(x_0\beta + \epsilon_i)$$
$$+ \frac{1}{n}\sum_{i=1}^{n} \delta_i \times h'(x_0\beta + \epsilon_i + \theta_i\delta_i). \quad (4.1)$$

By the strong law of large numbers, the first term on the right side of (4.1) is strongly consistent for the untransformed scale expectation $E\ Y_0$. It remains to show that the second term is stochastically small in some sense.

By the Cauchy-Schwarz inequality, the square of the second term in (4.1) is bounded from above by the product

$$\frac{1}{n}\sum_{i=1}^{n} \delta_i^2 \times \frac{1}{n}\sum_{i=1}^{n} [h'(x_0\beta + \epsilon_i + \theta_i\delta_i)]^2. \quad (4.2)$$

Lemma 1. Assume that (i) the retransformation h is continuously differentiable, (ii) X contains the intercept, and (iii) $X'X/n \to \Sigma$ positive definite, then

$$\frac{1}{n}\sum_{i=1}^{n} \delta_i^2 = \frac{1}{n}\sum_{i=1}^{n} [(x_0 - x_i)(X'X)^{-1}X'\epsilon]^2$$
$$= O_p(n^{-1}). \ \|$$

(The proof is straightforward and can be found in Duan et al. 1982, Appendix B, Addendum I.)

Assumption (iii) in the lemma is much stronger than we need. For most purposes in this article, it is sufficient to assume that $x_0(X'X/n)^{-1}x_0'$ is bounded. Nevertheless, the present assumption is satisfied for many problems—for example, when the covariates x_i are sampled randomly from a fixed parent population.

It follows from Lemma 1 that we can choose M large enough such that, for n large enough, the inequality

$$\sum_{i=1}^{n} \delta_i^2 < M^2 \quad (4.3)$$

holds with probability arbitrarily close to one. When (4.3) holds, we have

$$|\delta_i| < M \qquad i = 1, \ldots, n,$$
$$|h'(x_0\beta + \epsilon_i + \theta_i\delta_i)| \le \sup_{|t|\le M} |h'(x_0\beta + \epsilon_i + t)|;$$

thus

$$\frac{1}{n}\sum_{i=1}^{n} [h'(x_0\beta + \epsilon_i + \theta_i\delta_i)]^2$$
$$\le \frac{1}{n}\sum_{i=1}^{n} \sup_{|t|\le M} [h'(x_0\beta + \epsilon_i + t)]^2. \quad (4.4)$$

By the strong law of large numbers, the right side of (4.4) converges almost surely to

$$E \sup_{|t|\le M} [h'(x_0\beta + \epsilon + t)]^2 \quad (4.5)$$

if the expectation is finite.

To summarize, if the expectation (4.5) is finite for all $M > 0$, the second factor in (4.2) is bounded from above, with probability arbitrarily close to one, by a sequence of random variables that converge almost surely to a finite constant. In other words, the second factor in (4.2) is stochastically bounded. Thus we have proved the main result in this article.

Theorem 1. Assume (i) the retransformation h is continuously differentiable, (ii) X contains the intercept, (iii) $X'X/n \to \Sigma$ positive definite, and (iv) the expectation (4.5) is finite for all $M > 0$. Then the smearing estimate (3.1) is weakly consistent.

For most popular transformations, the supremum in (4.5) can be evaluated at the endpoints. For example, if $|h'|$ is monotonic, we have

$$E \sup_{|t|\le M} [h'(x_0\beta + \epsilon + t)]^2$$
$$\le E\,[h'(x_0\beta + \epsilon + M)]^2$$
$$+ E\,[h'(x_0\beta + \epsilon - M)]^2.$$

The moment condition (iv) in Theorem 1 can then be replaced by (v) $E\,[h'(c + \epsilon)]^2 < +\infty$ for all c, which is usually easy to check under hypothesized true error distribution. For example, for the power transformations

$$\eta = g(Y) = Y^{\alpha},\ \alpha \ne 0,$$
$$Y = h(\eta) = \eta^{1/\alpha},$$

the desired moment condition is that $E(c + \epsilon)^{2[(1/\alpha)-1]} < +\infty$ for all c, which is satisfied for the normal error distribution if $0 < \alpha < 1$. For the logarithmic transformation $\eta = \log(Y)$, $Y = \exp(\eta)$, the desired moment condition reduces to $E \exp(2\epsilon) < +\infty$, which is satisfied for the normal error distribution.

5. EFFICIENCY OF THE SMEARING ESTIMATE FOR LOGNORMAL LINEAR MODELS

If the error distribution is indeed normal, both the normal theory estimates and the smearing estimate are consistent, but the normal theory estimates can be more efficient. In this section we examine the loss of efficiency of the smearing estimate relative to normal theory estimates when the error distribution is indeed normal. For simplicity, we consider only the logarithmic transformation in this section.

The smearing estimate (SE), as defined in Section 3, is $\exp(x_0\hat{\beta}) \times n^{-1}\sum \exp(\hat{\epsilon}_i)$, where $\hat{\beta}$ are the least squares regression coefficients and $\hat{\epsilon}_i = \eta_i - x_i\hat{\beta}$ are the estimated residuals. We consider three normal theory estimates:

1. The naive estimate (NE), $\exp(x_0\hat{\beta} + \hat{\sigma}^2/2)$, where $\hat{\sigma}^2$ is the mean square for error;

2. The uniformly minimum variance unbiased estimate (MVUE),

$$\exp(x_0\hat{\beta})\, g_m\,[(m + 1)(1 - v_0)\hat{\sigma}^2/2m],$$

where m = degrees of freedom of $\hat{\sigma}^2$, $v_0 = x_0(X'X)^{-1}x_0'$, and

$$g_m(t) = \sum_{k=0}^{\infty} \frac{m^k(m + 2k)}{m(m + 2)\cdots(m + 2k)} \left(\frac{m}{m + 1}\right)^k \frac{t^k}{k!}.$$

3. Meulenberg's (1965) estimate (MEUL), $\exp[x_0\hat{\beta} + (1 - v_0)\hat{\sigma}^2/2]$.

The proof of the following theorem is straightforward and can be found in Duan et al. (1982, Appendix B).

Theorem 2. Assume (i) $X'X/n \rightarrow \Sigma$ positive definite, (ii) X contains the intercept, and (iii) $\epsilon \sim N(0, \sigma^2)$. Then

$$n\mathrm{var(SE)} \rightarrow [(x_0\Sigma^{-1}x_0')\sigma^2 + \exp(\sigma^2) - 1 - \sigma^2] \times \exp(2x_0\beta + \sigma^2),$$

$$n\mathrm{var(NE)} \rightarrow [(x_0\Sigma^{-1}x_0')\sigma^2 + \sigma^4/2] \times \exp(2x_0\beta + \sigma^2).$$

The asymptotic relative efficiency of the smearing estimate to the naive normal theory estimate is therefore

$$\frac{(x_0\Sigma^{-1}x_0')\sigma^2 + \sigma^4/2}{(x_0\Sigma^{-1}x_0')\sigma^2 + [\exp(\sigma^2) - 1 - \sigma^2]}. \tag{5.1}$$

Remarks.

1. Note that $\sigma^4/2$ is the leading term in Taylor's expansion of $\exp(\sigma^2) - 1 - \sigma^2$.
2. Using either Mehran's (1973) or Bradu and Mundlak's (1970) exact variance formula, one can show that the asymptotic variance of MVUE is the same as that of NE given in Theorem 2.
3. The proof in Duan et al. (1982, Appendix B, Addendum II, pp. 99–102) can be modified to show that the asymptotic variance of Meulenberg's estimate (MEUL) is the same as that of NE given in Theorem 2.
4. For the one population lognormal model with no covariates, the smearing estimate is the sample mean, and $x_0\Sigma^{-1}x_0' = 1$. Theorem 2 is then equivalent to the asymptotic case in Mehran's (1973) comparison of the sample mean and the normal theory estimate MVUE. The asymptotic variance formula for NE is the first-order term in Finney's (1941) approximation for this special case.

The relative efficiency depends on both σ^2 and $x_0\Sigma^{-1}x_0'$. If x_0 is sampled randomly from the same population as x_i's, we have

$$\begin{aligned} \mathrm{E}\, x_0\Sigma^{-1}x_0' &= \mathrm{tr}\Sigma^{-1}\mathrm{E}\, x_0'x_0 \\ &= \mathrm{tr}\Sigma^{-1}\Sigma \quad (\Sigma = \mathrm{E}\, x'x) \\ &= \mathrm{rank}(X); \end{aligned}$$

thus $x_0\Sigma^{-1}x_0'$ is of the same order as rank(X). Table 1 contains the relative efficiency for a wide range of values of σ^2 and $x_0\Sigma^{-1}x_0'$. For σ^2 near or less than one, the relative efficiency is very high. Mehran (1973) also noted that the sample mean performed "surprisingly well" relative to MVUE for the one population lognormal model. The column "$x_0\Sigma^{-1}x_0' = 1$" in Table 1 is equivalent to the column "$n = \infty$" in Mehran's table.

Table 1. Relative Efficiency of the Smearing Estimate to the Normal Theory Estimate When the Normality Assumption is Satisfied

	$x_0\Sigma^{-1}x_0'$ ($\approx$ *rank X*)				
σ^2	*1*	*2*	*3*	*10*	*20*
.10	1.00	1.00	1.00	1.00	1.00
.50	.96	.98	.99	1.00	1.00
1.00	.87	.92	.94	.98	.99
2.00	.63	.72	.77	.90	.95
3.00	.39	.48	.54	.75	.85

For large σ^2, the relative efficiency drops drastically. Under the assumed model, the untransformed scale responses follow a lognormal distribution, with σ^2 being the shape parameter: large σ^2 indicates large skewness.

For most empirical problems, the values of σ^2 are likely to belong to the range for which the smearing estimate has very high relative efficiency compared with the normal theory estimates. Goldberger (1968) states that "a casual survey of empirical work suggests that σ^2, the logarithmic disturbance, is unlikely to exceed 0.5." Mincer (1974, p. 101) tabulated the variances of log annual (1959) earning within age × education groups. Most of the variances are near or below .5, except that several groups of higher-educated people have variances as high as .933, which are still within the range for which the smearing estimate has very high relative efficiency.

The range of σ^2 might, however, depend on the field of application. Ott, Mage, and Randecker (1979) analyzed carbon monoxide concentration data from 11 U.S. cities. The variances of the log concentrations in some cities are rather high: 1.94 for Phoenix, Arizona; 1.42 for Barstow, California; and 1.27 for Denver, Colorado. Some of these values are in the range for which the smearing estimate does not have very high relative efficiency. For example, the relative efficiency for Phoenix is only 64 percent ($\sigma^2 = 1.94$, $x_0\Sigma^{-1}x_0' = 1$ for the one population lognormal model).

For large σ^2, while the normal theory estimates are substantially more efficient than the smearing estimate when the normal assumption is true, they can also be more sensitive to departures from normality. As an illustration, we will consider again the example used in Section 2:

$$\begin{aligned} &\epsilon \sim N(0, .95\sigma^2) \quad \text{with probability } .995, \\ &\epsilon \sim N(0, 10.95\sigma^2) \quad \text{with probability } .005, \\ &(\mathrm{var}\ \epsilon = \sigma^2). \end{aligned}$$

Table 2 provides the asymptotic relative bias of the normal theory estimate. Under the assumed model, the

Table 2. Asymptotic Relative Bias of the Normal Theory Estimate Under the Mixture Model

σ^2	*Asymptotic Relative Bias (%)*
.50	−4
.75	−16
1.00	−41
1.50	−90
2.00	−99
3.00	−100

asymptotic relative bias is

$$\frac{\exp(x_0\beta + \sigma^2/2)}{\exp(x_0\beta) \cdot [.995 \exp(.475\sigma^2) + .005 \exp(5.475\sigma^2)]} - 1.$$

The asymptotic relative bias increases with σ^2; for σ^2 near or larger than one, the normal theory estimates are severely biased.

6. HEALTH INSURANCE STUDY EXPENDITURE ANALYSIS

To illustrate the use of the smearing estimate, we consider a real-world problem, predicting individual medical expenditures on the Rand Health Insurance Study (HIS). (Rubin 1983 and Morris 1983 address similar problems in a broader context.) The study is a longitudinal social experiment designed to study, among other things, how different health insurance policies affect the demand for health care. A random sample of 2,756 families from six sites across the U.S. are assigned to 14 different insurance plans that vary the amount of cost sharing. The study reimburses their insurance claims, thereby obtaining a measure of their demand for health care. Further details on the study can be found in Newhouse et al. (1981).

Duan et al. (1982) have proposed a model for individual health expenditures on HIS, one part of which models the annual expenditure of individuals with positive ambulatory expenditure and no inpatient expenditure in a given year as a linear regression model on the log scale:

$$\eta_i = \log Y_i = \mathbf{x}_i\boldsymbol{\beta} + \epsilon_i, \qquad (6.1)$$

where Y_i denotes the annual expenditure. The explanatory variables $\mathbf{x}$ include five different levels of experimental cost sharing, demographic characteristics, preexperimental use of medical care, and the individual's self-perception of his health. The sample is partitioned into nine subsamples, each consisting of a specific year from a specific site. Equation (6.1) is fitted to the subsamples separately. The sample sizes in the nine subsamples range from 501 to 857; the total sample size is 6,479. The number of explanatory variables range from 24 to 27, depending on the subsample; some explanatory variables are not available or are not applicable in certain sites. Further details on the sample and the analysis can be found in Duan et al. (1982).

The error distribution in (6.1) is fairly close to a normal distribution, but it is slightly skewed towards the lower extreme. (Skewness = −.37, Kolmogorov-Smirnov statistic = .034.) Apparently both the normal theory estimate and the smearing estimate are plausible candidates for predicting the untransformed scale expectation.

As was noted in the previous section, the relative efficiency of the smearing estimate depends on the error variance $\sigma^2 = \text{var}(\epsilon)$. The estimated σ^2's for the nine subsamples range from 1.03 to 1.34, and fall mostly in the range in which the smearing estimate is fairly efficient under the normal assumption.

To compare the prediction performance of various models, Duan et al. (1982) developed a cross-validation-type technique. They split the sample randomly into two parts, the training subsample and test subsample; the various models are fitted on the training subsample, and then used to form predictions for all individuals in the test subsample. The prediction for each individual is then compared with that individual's response actually observed. The average squared prediction error (ASPE) is then computed for each model:

$$\text{ASPE} = \sum (\hat{\text{E}}\, Y_k - Y_k)^2,$$

where the index k runs through the m individuals in the test subsample. The ASPE's for different models can be compared directly, the model producing the smaller ASPE being the better one.

For the HIS data, the smearing estimate has smaller ASPE (37429) than the normal theory estimates. For the naive estimate, ASPE = 38908; for MVUE, ASPE = 38081; and for MEUL, ASPE = 38093. The differences might appear to be only a minor fraction of ASPE. However, it should be noted that ASPE is a combination of estimation error and measurement error. A minor improvement in terms of ASPE might actually be a major improvement in terms of estimation error. (The methodology presented here does not distinguish estimation error from measurement error, though.)

An alternative procedure, the subpopulation sign test, also proposed in Duan et al. (1982), divides the test sample into subpopulations. The ASPE's for each model are computed on each subpopulation. When comparing two models, the number of subpopulations on which each model has smaller ASPE than the other is calculated. The model that wins a larger number of subpopulations is judged to be the better model.

The HIS data are naturally partitioned into 43 subpopulations, each consisting of a distinct combination of site, year, and experimental plan. Under a null hypothesis of no differences between the estimation techniques, the number of winning subpopulations follow a binomial distribution with sample size 43 and null probability .5. The .05 two-sided rejection region is $\{\# < 15\} + \{\# > 28\}$. The smearing estimate wins 29 of the subpopulations when compared with the naive estimate. Compared with MVUE or MEUL, the smearing estimate wins 28 subpopulations.

It should also be noted that the average bias,

$$\text{Bias} = (\text{average of individual predictions}) - (\text{average of actual observations}),$$

and relative bias,

$$\text{Relative bias} = \text{Bias}/(\text{average of actual observations}),$$

on the test sample are much smaller for the smearing estimates (bias = \$7.00, relative bias = 4.3 percent) than the normal theory estimates. For the naive estimate, bias = \$22.00, relative bias = 13.7 percent; for MVUE, bias = \$13.70, relative bias = 8.5 percent; for MEUL, bias = \$13.90, relative bias = 8.6 percent.

Based on ASPE, on the subpopulation sign test, and on the bias, we conclude that the smearing estimate is the more appropriate procedure to use in this case.

[*Received December 1981. Revised February 1983.*]

REFERENCES

BRADU, D., and MUNDLAK, Y. (1970), "Estimation in Lognormal Linear Models," *Journal of the American Statistical Association*, 65, 198–211.

COX, D.R., and HINKLEY, D.V. (1968), "A Note on the Efficiency of Least Squares Estimates," *Journal of the Royal Statistical Society*, Ser. B, 30, 284–289.

DUAN, NAIHUA, MANNING, WILLARD G., MORRIS, CARL N., and NEWHOUSE, JOSEPH P. (1983), *A Comparison of Alternative Models for the Demand for Medical Care*, R-2754-HHS, Santa Monica, Cal.: The Rand Corporation. (Also in *Journal of Business & Economic Statistics*, 1, 115–126.

EBBELER, D.H. (1973), "A Note on Large-Sample Approximation in Lognormal Linear Models," *Journal of the American Statistical Association*, 68, 231.

EFRON, B. (1979), "Bootstrap Methods: Another Look at the Jackknife," *The Annals of Statistics*, 7, 1, 1–26.

EVANS, I.G., and SHABAN, S.A. (1974), "A Note on Estimation in Lognormal Models," *Journal of the American Statistical Association*, 69, 779–781.

FINNEY, D.J. (1941), "On the Distribution of a Variate Whose Logarithm is Normally Distributed," Supplement to *Journal of the Royal Statistical Society*, 7, 155–161.

GOLDBERGER, A.S. (1968), "The Interpretation and Estimation of Cobb-Douglas Functions," *Econometrica*, 35, 464–472.

MEHRAN, F. (1973), "Variance of the MVUE for the Lognormal Mean," *Journal of the American Statistical Association*, 68, 726–727.

MEULENBERG, M.T.G. (1965), "On the Estimation of an Exponential Function," *Econometrica*, 33, 863–868.

MINCER, J. (1974), *Schooling, Experience, and Earning*, New York: Columbia University Press.

MORRIS, CARL N. (1983), Discussion of "A Case Study of Bayesian/Likelihood Methods of Inference: Estimating the Total in a Finite Population Using Transformations to Normality," by D. Rubin, to appear in the *Proceedings of the Conference on Scientific Inference, Data Analysis, and Robustness*, New York: Academy Press (in press).

NEWHOUSE, JOSEPH P., et al. (1981), "Some Interim Results from a Controlled Trial of Cost Sharing in Health Insurance," *New England Journal of Medicine*, 305, 1501–1507.

NEYMAN, JERZY, and SCOTT, ELIZABETH (1960), "Correction for Bias Introduced by a Transformation of Variables," *Annals of Mathematical Statistics*, 31, 643–655.

OTT, W.R., MAGE, D.T., and RANDECKER, V.S. (1979), "Testing the Validity of the Lognormal Probability Model: Computer Analysis of Carbon Monoxide Data from U.S. Cities," EPA-600/4-79-040, U.S. Environmental Protection Agency, Washington, D.C.

RUBIN, D. (1983), "A Case Study of Bayesian/Likelihood Methods of Inference: Estimating the Total in a Finite Population Using Transformations to Normality," to appear in the *Proceedings of the Conference on Scientific Inference, Data Analysis, and Robustness*, New York: Academy Press (in press).

SHIMIZU, K., and IWASE, K. (1981), "Uniformly Minimum Variance Unbiased Estimation in Lognormal and Related Distributions," *Communications in Statistics*, A, 11, 687–697.