DESIGN ASPECTS OF REGRESSION BASED RATIO ESTIMATION

by

Gary C. McDonald
Mathematics Department
General Motors Research Labs
Warren, MI 48090–9055

William J. Studden*
Purdue University
Department of Statistics
Mathematical Science Bldg
West Lafayette, IN 47907

Technical Report # 86-53

Department of Statistics
Purdue University

November, 1986
Revised September, 1988

1. INTRODUCTION

This article will focus on regression based ratio estimators which are formed by taking the ratio of mean value estimates, based on a linear regression function, at two distinct points. Such estimators arise in the context of assessing the percentage change of the mean value of a response variable over a specified region in the space of regressor variables, a quantity which is nonlinear in the regression parameters. An important application, which motivated this work, arises in a regulatory context to assess the automobile emissions pattern as a function of accumulated mileage. This assessment is capsulized in the so-called "deterioration factor" which is the ratio of two points obtained from a least squares line fitted to emission-mileage data. McDonald (1981, 1988) discusses this problem, constructs confidence regions for the deterioration factor using Fieller's method (1954), and investigates experimental design issues (i.e., at which mileages should emission measurements be taken) so as to minimize the length of the resultant confidence interval. Several approximations to an optimal design criterion were derived. Within a restricted class of designs, optimal designs were derived which quantified the tradeoff between reduced mileage accumulation and increased number of emissions tests (i.e., number of data points).

In this article we provide further results on the statistical design issues, arising within the ratio context, along a number of important lines. The problem is now specified as a general multiple linear model thus permitting regression functions which are, perhaps, quadratic or even spline-like. The length of the confidence interval is related to the asymptotic variance of the regression based ratio estimator and thus is used as a basis for constructing optimal experimental designs in the case of a simple linear model. This methodology is extended further to the case of a potentially changing regression slope; i.e., inclusion of a spline knot in the design region.

Buonaccorsi and Iyer (1984) consider a related problem—that of estimating the point at which a regression function, quadratic in one variable, achieves a maximum or minimum. If $E(y) = \beta_0 + \beta_1 x + \beta_2 x^2$, then the parameter of interest is $\rho = -\beta_1/2\beta_2$ and the authors consider several methods of obtaining a confidence region for ρ along with optimal design implications. In a recent article (1986) those same authors consider estimating a ratio of two linear combinations of the vector of parameters in the general linear model and optimizing the design with respect to the asymptotic variance of the estimator. In this article we consider a different application and extend the optimal design considerations to include spline type models and to include applications where only a portion of the design points are free to be placed in an optimal fashion.

2. MATHEMATICAL FORMULATION

We assume a general linear model of the form

$$y_i = \mathbf{f}'(x_i)\beta + e_i, \qquad i = 1, \ldots, n, \tag{2.1}$$

where $\mathbf{y}' = (y_1, \ldots, y_n)$ is an $1 \times n$ vector of observations; β is a $p \times 1$ vector of regression coefficients $\mathbf{f}(x_i)$ is a $p \times 1$ vector of regression functions evaluated at $x_i \in [a, b]$; and $\mathbf{e}' = (e_1, \ldots, e_n)$ is an $1 \times n$ vector of error terms assumed to be independent and identically distributed with a normal distribution having mean zero and variance σ^2. The least squares estimate of β is given by

$$\mathbf{b} = (X'X)^{-1}X'y \tag{2.2}$$

where X is the $n \times p$ design matrix given by

$$X = \begin{pmatrix} f_1(x_1) & f_2(x_1) & \ldots & f_p(x_1) \\ \vdots & & & \\ f_1(x_n) & f_2(x_n) & \ldots & f_p(x_n) \end{pmatrix}. \tag{2.3}$$

It will be convenient, for later purposes, to let $M = (X'X)/n$ denote the information matrix per observation. The variance-covariance matrix of the estimates $\mathbf{b}$ is then given by

$$\begin{aligned} \mathrm{Cov}(\mathbf{b}) &= \sigma^2(X'X)^{-1} \\ &= (\sigma^2/n)M^{-1}. \end{aligned} \tag{2.4}$$

The variance of the error term is estimated by the usual unbiased estimator based on $n - p$ degrees of freedom (df); that is, by

$$s^2 = (n-p)^{-1}(\mathbf{y}'\mathbf{y} - \mathbf{b}'X'\mathbf{y}) \tag{2.5}$$

Now let $\mu_i = E(y(x_i)) = \mathbf{f}'(x_i)\beta$ be the mean value at two specified points with, say, $x_1 > x_2$. The usual estimates $\hat{\mu}_i = \mathbf{f}'(x_i)\mathbf{b}$, $i = 1, 2$, have a 2×2 variance-covariance matrix given by

$$\mathrm{Cov}(\hat{\mu}_1, \hat{\mu}_2) = (\sigma^2/n)V, \tag{2.6}$$

where

$$V = (v_{ij}) = \begin{pmatrix} \mathbf{f}'(x_1) \\ \mathbf{f}'(x_2) \end{pmatrix} M^{-1}(\mathbf{f}(x_1), \mathbf{f}(x_2)). \tag{2.7}$$

If $\mu_2 \neq 0$, the statement $\mu_1/\mu_2 = \theta$ is equivalent to $\mathbf{f}'(x_1)\beta - \theta\mathbf{f}'(x_2)\beta = 0$. Thus a hypothesis of the form

$$H\colon\ \mu_1/\mu_2 = \theta \tag{2.8}$$

can be written as a special case of the general linear hypothesis (Searle 1971, Chap. 3); that is,

$$H\colon \lambda_\theta'\beta = 0, \tag{2.9}$$

where

$$\lambda_\theta' = \mathbf{f}'(x_1) - \theta\mathbf{f}'(x_2). \tag{2.10}$$

In order to place a confidence interval on the quantity μ_1/μ_2, the F-test for the hypothesis H is inverted. All θ values that lead to acceptance of H (versus an alternative of inequality) are placed in the confidence set. The critical region for the F-test for H is given by

$$\text{Reject } H \text{ iff } Q \geq s^2 F_{1,n-p;\alpha}, \tag{2.11}$$

where

$$Q = (\lambda_\theta'\mathbf{b})^2[\lambda_\theta'(X'X)^{-1}\lambda_\theta]^{-1}, \tag{2.12}$$

and $F_{1,n-p;\alpha} \equiv F$ is the upper α^{th} percentage point of a central F-variate with 1 and $n-p$ df. The quantity s^2 is given by equation (2.5). Thus, all values of θ for which

$$Q < s^2 F \tag{2.13}$$

are included in a $100(1-\alpha)$ percent confidence region (CR) for the ratio μ_1/μ_2.

The inequality (2.13) can be expressed conveniently in a quadratic form. Let

$$L = \mathbf{b}\mathbf{b}' - (s^2/n)FM^{-1}, \tag{2.14}$$

and

$$\begin{aligned} C = (c_{ij}) &= \begin{pmatrix} \mathbf{f}'(x_1) \\ \mathbf{f}'(x_2) \end{pmatrix} L(\mathbf{f}(x_1), \mathbf{f}(x_2)) \\ &= \begin{pmatrix} \hat{\mu}_1 \\ \hat{\mu}_2 \end{pmatrix} (\hat{\mu}_1, \hat{\mu}_2) - (s^2/n)FV. \end{aligned} \tag{2.15}$$

Then (2.13) reduces to

$$(1, -\theta)C\begin{pmatrix} 1 \\ -\theta \end{pmatrix} < 0,$$

or

$$\theta^2 c_{22} - 2\theta c_{12} + c_{11} < 0. \tag{2.16}$$

Thus θ satisfying (2.16) would comprise the $100(1-\alpha)\%$ CR for μ_1/μ_2. If $c_{22} > 0$, then the region can be expressed as an interval (CI)

$$\theta \ \in \ [c_{12}/c_{22} \pm (c_{12}^2 - c_{11}c_{22})^{1/2}/c_{22}]. \tag{2.17}$$

The inequality $c_{22} > 0$ is

$$c_{22} = \hat{\mu}_2^2 - (s^2/n)Fv_{22} > 0,$$

or

$$\hat{\mu}_2^2[(s^2/n)F]^{-1} > v_{22}.$$

This will be true if either the signal to noise ratio at x_2 is sufficiently large or n is large.

Further discussion of the form of the region determined from the parabola on the left-hand side of (2.16) is given by Kendall and Stuart (1973, pp. 130–132).

The center of the CI on θ can be written as

$$\begin{aligned} c_{12}/c_{22} &= [\hat{\mu}_1\hat{\mu}_2 - (s^2/n)Fv_{12}]/[\hat{\mu}_2^2 - (s^2/n)Fv_{22}] \\ &\to \mu_1/\mu_2 \quad \text{as} \quad n \to \infty. \end{aligned} \tag{2.18}$$

This limiting value, as expected, is simply the ratio of the response estimates at x_1 and x_2. In general, the center value c_{12}/c_{22} will not be $\hat{\mu}_1/\hat{\mu}_2$. In fact

$$c_{12}/c_{22} > \hat{\mu}_1/\hat{\mu}_2$$

if and only if

$$\hat{\mu}_1/\hat{\mu}_2 > v_{12}/v_{22}.$$

In some applications the response $b_0 + b_1x$ is increasing so the left hand side is greater than one while the right hand side, expressible as

$$v_{12}/v_{22} = \frac{v_{12}}{\sqrt{v_{11}}\sqrt{v_{22}}} \cdot \frac{\sqrt{v_{11}}}{\sqrt{v_{22}}},$$

will be less than one if v_{11} is no greater than v_{22}.

The length of the CI is given by $2(c_{12}^2 - c_{11}c_{22})^{1/2}/c_{22}$. The matrix C is given in (2.15). This expression is sufficiently complicated to warrant considering a related normalization factor which can be used for design purposes. We choose the variance of the limiting distribution (see Appendix A) to serve as such a factor and denote it by Var; i.e.,

$$\text{Var }(\hat{\mu}_1/\hat{\mu}_2) = (\sigma^2/n)\mu_2^{-4}(\mu_2, -\mu_1)V\begin{pmatrix} \mu_2 \\ -\mu_1 \end{pmatrix}. \tag{2.19}$$

Additionally, McDonald (1988) gives conditions under which a design minimizing the expected length of the confidence interval could be determined, approximately, by choosing a design to maximize $|X'X|/[\beta'X'X\beta + \sigma^2(1-F)]$. For large n this would be the design minimizing the expression given in (2.19). We will refer to Var $(\hat{\mu}_1/\hat{\mu}_2)$ as the asymptotic variance.

3. SPECIAL CASE: SIMPLE LINEAR REGRESSION

In the special case where $E(y) = \beta_0 + \beta_1 x$ then $\mu_i = \beta_0 + \beta_1 x_i, i = 1, 2$, and equation (2.19) can be simplified. Using the expression (2.7) for V we find that

$$\text{Var}(\hat{\mu}_1/\hat{\mu}_2) = (\sigma^2/n)(x_1 - x_2)^2\mu_2^{-4}(-\beta_1, \beta_0)M^{-1}\begin{pmatrix} -\beta_1 \\ \beta_0 \end{pmatrix}. \tag{3.1}$$

An alternate expression, which for some purposes is more convenient, can be derived. As noted in (2.6), $(\sigma^2/n)V$ is the variance-covariance matrix of the estimates $\hat{\mu}_1$ and $\hat{\mu}_2$. In this case the matrix V is invariant under a change of basis of the regression functions $(1, x)$. Instead of 1 and x we use

$$\ell_a(x) = (b-x)/(b-a)$$

and

$$\ell_b(x) = (x-a)/(b-a), \tag{3.2}$$

where the design variable $x \in [a, b]$. In terms of these functions the response can be expressed as

$$\beta_0 + \beta_1 x = \mu_a \ell_a(x) + \mu_b \ell_b(x),$$

where μ_a and μ_b are the expected values of the response at a and b respectively. Using (3.2) the expression (2.19) can be written as

$$\mathrm{Var}(\hat{\mu}_1/\hat{\mu}_2) = (\sigma^2/n)[(x_1-x_2)/\mu_2^2(b-a)]^2(-\mu_b, \mu_a)M_\ell^{-1}\begin{pmatrix}-\mu_b\\ \mu_a\end{pmatrix}, \tag{3.3}$$

where M_ℓ is calculated in the same manner as $M = (X'X)/n$ using the regressors in (3.2).

Observations (i.e., design points in $[a, b]$), upon which the regression function is estimated, can be chosen to minimize the expression as given in (3.1) or (3.3).

Lemma 3.1. The minimization of (3.1) or (3.3) is achieved by placing the design points at a and b proportional to μ_b and μ_a respectively. That is, the number of observations N_a and N_b taken at a and b, respectively, is given by

$$N_a = n\mu_b/(\mu_a + \mu_b),$$

and (3.4)

$$N_b = n\mu_a/(\mu_a + \mu_b).$$

The minimum value is

$$\mathrm{Var}^*\ (\hat{\mu}_1/\hat{\mu}_2) = (\sigma^2/n)[(x_1-x_2)(\mu_a+\mu_b)/\mu_2^2(b-a)]^2. \tag{3.5}$$

Proof: This result can be derived from (3.1) using Elfving's Theorem; see, for example, Elfving (1952) or Karlin and Studden (1966). For our special case we can use the fact that for simple linear regression it is always best to choose all the observations at the interval endpoints, a and b. If this is done the matrix M_ℓ in (3.3) is diagonal with diagonal elements equal to the proportion of observations at a and b. The minimization is then easy to carry out to obtain (3.4) and (3.5).

Remark: The result in Lemma 3.1 applies to the asymptotic variance and is locally optimum in the sense that it depends on μ_a and μ_b or β_0 and β_1. Calculations on actual data (next section) show that the results are accurate even for sample sizes of 10. The local dependence is useful since prior information on the unknown parameters may be available or a sequential adaptive type procedure may be used. In addition the minimum value in (3.5) serves as a benchmark as to what could be attained.

Generally conditions will warrant or will require some observations to be taken in the middle of the design interval. In this case the information matrix M in (3.1) or (3.3) can be written in the form

$$\gamma_1 M_1 + \gamma_2 M_2 \tag{3.6}$$

where γ_1 is the proportion of the observations required or specified with corresponding information matrix M_1. The proportion $\gamma_2 \geq 0$ are free for design purposes so that now (3.1) is minimized with respect to M_2.

In the following we assume M_1 and M_2 are calculated using the regressors $\ell_a(x)$ and $\ell_b(x)$ in (3.2). Let the elements of M_1 be denoted by

$$M_1 = \begin{pmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{pmatrix}. \tag{3.7}$$

In minimizing the expressions (3.1) or (3.3) with respect to M_2 one would expect the resulting design to be close, in some sense, to the overall optimum given in Lemma 3.1. Let

$$p_a = \frac{\gamma_2\mu_b + \gamma_1[\mu_b(m_{22} - m_{12}) + \mu_a(m_{12} - m_{11})]}{\gamma_2(\mu_a + \mu_b)}. \tag{3.8}$$

The optimal design in this case is given in the following lemma. The proof is given in Appendix B. An example illustrating the use of Lemma 3.1 and 3.2 is given in Section 4.

Lemma 3.2. The minimization of (3.1) or (3.3), with respect to the class of information matrices given in (3.6), is achieved by placing the remaining proportion γ_2 of design points at the endpoints a and b in the proportions p_a and $1 - p_a$ where p_a is given by (3.8). If $p_a \geq 1$ ($p_a \leq 0$) then all the observations are taken at a (at b).

Remark: The quantity p_a in (3.8) is the proportion of the remaining observations assigned to the endpoint a. Note that if $\gamma_1 = 0$ (and $\gamma_2 = 1$) so that all of our observations are available for design purposes, then $p_a = \mu_b/(\mu_b + \mu_a)$. This is the solution given in Lemma 3.1.

Suppose that the prescribed proportion of observations γ_1 are assigned to $x_1, x_2, \ldots, x_k$ in proportion $\xi_1, \xi_2, \ldots, \xi_k$. The quantities m_{11}, m_{22}, m_{12} are then given by

$$\Sigma\xi_i\ell_a^2(x_i), \quad \Sigma\xi_i\ell_b^2(x_i), \quad \Sigma\xi_i\ell_a(x_i)\ell_b(x_i). \tag{3.9}$$

An inspection of $\ell_a(x)$ and $\ell_b(x)$ in (3.2) shows that the quantities m_{11}, m_{22}, and m_{12} give a measure of the proportion of observations near a, near b, and near the middle respectively. Note that $m_{12} = 0$ only if all the observations are at a and b. The relative sizes of these quantities (as well as μ_a and μ_b) will determine whether p_a is greater or less than $\mu_b/(\mu_b + \mu_a)$.

4. AN AUTOMOTIVE EMISSIONS EXAMPLE

McDonald (1981) gives automotive emissions data on hydrocarbon (HC), carbon monoxide (CO), and nitrogen oxides (NO_x) as a function of mileage accumulated on the vehicle. For illustrative purposes we consider here the HC emissions:

Miles (in 1000's)	HC (gm/mi)
5.133	0.265
10.124	0.278
15.060	0.282
19.946	0.286
24.899	0.310
29.792	0.333
29.877	0.343
35.011	0.335
39.878	0.311
44.862	0.345
49.795	0.319

The regression estimates are $b_0 = .2659$, $b_1 = .00158$, $s = .01775$, and $R^2 = .644$. Choosing, for this application, $x_1 = 50$ and $x_2 = 4$ we have $\hat{\mu}_1 = .3451$ and $\hat{\mu}_2 = .2722$. This selection of x_1 and x_2, along with the specification of the observation interval, is motivated in the McDonald (1981) reference. Thus $\hat{\mu}_1/\hat{\mu}_2 = 1.268$ and, using (2.17), the upper and lower limits for a 95% CI for μ_1/μ_2 are 1.454 and 1.109 respectively.

The expression in (3.1) can be used to approximate a CI for the ratio. Based on the eleven observations the resulting value is $\text{Var}(\hat{\mu}_1/\hat{\mu}_2) = .00574$. The corresponding approximate interval estimate, i.e., $c_{12}/c_{22} \pm [F \cdot \text{ var } (\hat{\mu}_1/\hat{\mu}_2)]^{1/2}$ is from 1.110 to 1.453 which agrees remarkably well with 1.109 and 1.454.

Taking the observation interval to be from $a = 5$ to $b = 50$ Lemma 3.1 yields the optimal observation placement to be 56% of the observations at $a = 5$ and 44% at $b = 50$. For $n = 11$ the best allocation would be 6 observations at a and 5 observations at b. A larger regression slope would generally require more observations at the left endpoint, $a = 5$. Note that we should have $N_a < N_b$ if the slope β_1 is negative.

It is interesting to note that the locally optimal design given in (3.4) depends on the values μ_a and μ_b of the regression function at the endpoints a and b but not on the points x_1 and x_2. This is due to the fact that in minimizing the right side of (3.3) the design or choice of observation enters through M_ℓ and hence we minimize

$$(-\mu_b, \mu_a) M_\ell^{-1} \begin{pmatrix} -\mu_b \\ \mu_a \end{pmatrix}.$$

The asymptotic variance in (3.3) depends on x_1 and x_2 through the multiplier in front involving $(x_1 - x_2)^2$.

The minimal asymptotic variance, from (3.5), is approximately

$$\text{Var}^* \, (\hat{\mu}_1/\hat{\mu}_2) = .02297/n.$$

For $n = 11$ this gives a standard error of .0457. This compares with a value of .0758 for the original design. The design achieving the minimal asymptotic variance would thus yield a CI approximately 40% shorter than that based on the original design.

Using the same data we can illustrate the design considerations if observations are required at certain mileage values. Suppose for example that observations where required at $x = 5$, $x = 15$, and two observations where required at the maintenance value $x = 30$. Simple calculations from (3.9) show that $m_{11} = 1/2$, $m_{12} = m_{22} = 1/6$. If a total of 11 observations were allowed as before then $\gamma_1 = 4/11$ and $\gamma_2 = 7/11$. Inserting these values together with $\mu_a = .2738$ and $\mu_b = .3451$ into (3.8) we find that $p_a = .473$. In this case the remaining 7 observations might best be allocated by placing 3 at the left endpoint a and 4 at the right endpoint b. This allocation yields four observations at $x = 5$; one at $x = 15$; two at $x = 30$; and four at $x = 50$. For this a recalculation of the asymptotic variance in (3.3) gives the value

$$\text{Var}(\hat{\mu}_1/\hat{\mu}_2) = .0305/n.$$

For $n = 11$ this now gives a standard error of .0526. This value is, of course, between the value .0763 for the original design and .0457 for the optimum unrestricted design using Lemma 3.1.

5. LINEAR SPLINE REGRESSION

The automotive emissions applications are conducted with certain maintenance being performed at some point, roughly midway, between a and b. Allowing the possibility that the emission vs. mileage response function might change at the maintenance point, we could introduce a continuous segmented line model

$$E(y) = \beta_0 + \beta_1 x + \beta_2 (x - \xi)_+, \tag{5.1}$$

where $(x - \xi)_+ = x - \xi$ if $x > \xi$ and equals zero if $x \leq \xi$. This provides for a possible slope change at the point ξ, $a < \xi < b$.

To describe the corresponding design considerations, let μ_a, μ_ξ, μ_b denote the mean response values at $x = a, \xi$, and b respectively. Further let $\rho_a = (a - x_2)/(\xi - x_2)$ and $\rho_b = (x_1 - b)/(x_1 - \xi)$ where we assume only that $x_2 < \xi < x_1$. The following lemma gives the optimal design. The proof is sketched in Appendix C.

Lemma 5.1. For the model (5.1) the locally optimal design minimizing the variance (2.20) places observations at the three points a, ξ, and b proportional to the three quantities

$$w_1 = |\mu_b - \mu_\xi \rho_b|, \quad w_2 = |\mu_b \rho_a - \mu_a \rho_b|, \quad w_3 = |\mu_a - \mu_\xi \rho_a| \tag{5.2}$$

respectively. The optimal value of the variance is given by

$$\text{Var}^* \left(\hat{\mu}_1/\hat{\mu}_2\right) = \frac{\sigma^2}{n}\frac{1}{\mu_2^4}\rho^2 \left(\sum_{i=1}^{3} w_i\right)^2 \tag{5.3}$$

where $\rho = (\xi - x_2)(x_1 - \xi)/(\xi - a)(b - \xi)$.

As an example of Lemma 5.1 we applied it and the model (5.1) with $\xi = 29.792$ to the data in Section 4. The estimates of the parameters in (5.1) are $b_0 = .2434$, $b_1 = .002922$ and $b_2 = -.00316$ while $s = .01246$ and $R^2 = .8439$. The s value decreased slightly from its value of .01775 in the simple regression case. The corresponding estimates of μ_1 and μ_2 are $\hat{\mu}_1 = .3256$ and $\hat{\mu}_2 = .2550$. A little further calculation using (2.19) gives an asymptotic standard error (of the ratio estimate using the spline) of .0614. The corresponding 95% CI for μ_1/μ_2 is from 1.15 to 1.40. The interval in this case is somewhat smaller.

We can again calculate, using Lemma 5.1 the increase in accuracy if an optimal design had been used. The design in the spline case is also of a local nature, i.e., it depends on the parameters. The calculations below are based on the estimates of the parameters given above.

The quantities w_1, w_2, w_3 are easily seen to be .3256, .0126, and .2453 while $\rho = 1.0403$. Using $s = .01246$ and $\hat{\mu}_2 = .2550$ the optimal asymptotic variance in this case is

$$\text{Var}(\hat{\mu}_1/\hat{\mu}_2) = \frac{.0135}{n}$$

This value should be compared with $.02297/n$ for the simple regression case. For $n = 11$, the standard error is .0350, which, when compared with .0614 for the original design indicates a reduction of 43% in the length of the interval.

6. SUMMARY

An important subclass of regression problems is the estimation, both point and interval, of a ratio of mean values. Specifying statistical designs (i.e., the frequency and placement of points in the design space where observations are to be taken) is challenging and difficult due to the nonlinear nature of the function to be estimated. In this article we have focused, primarily, on the asymptotic variance of the ratio of the least squares estimates of mean values as a basis upon which to construct optimal designs. "Optimal" in this context refers to minimization of a normalization factor derived as the variance of the related asymptotic distribution.

In the case of a simple linear regression or a continuous segmented linear regression it is possible to obtain explicit optimal designs. As in estimating the regression slope, the optimal design places all the mass at the interval endpoints. The proportion of observations allocated to the individual endpoints depends on the regression response at those points and, hence, is locally optimal. Numerical calculations in a specific case (sample size 11) suggest that while such designs are based on asymptotic variances they are, in fact, quite good for exact interval expected length minimization based on a more complex approach. Additionally, optimal designs are derived for the important applications where some proportion of the observations are required to be taken in a specified fashion and the remaining proportion are free to be allocated for design purposes.

These optimal design results further serve to indicate how efficient other designs might be; e.g., those strategies used for the construction of automotive emission deterioration factors. In this application these statistical measures augment other goals of such a testing program which relate to hardware integrity. Such measures provide a scale upon which various testing programs (including regulatory) for establishing deterioration factors can be compared in terms of statistical estimation accuracy, number of required tests, and the associated accumulated mileage (which translates to calendar time).

REFERENCES

Buonaccorsi, J. P. and Iyer, H. K. (1984), "A Comparison of Confidence Regions and Designs in Estimation of a Ratio," *Commun. Statist.-Simula. Computa.,* **13**(6), 723–741.

Buonaccorsi, J. P. and Iyer, H. K. (1986), "Optimal Designs for Ratios of Linear Combinations in the General Linear Model," *JSPI* **13**(3), p.345–356.

Elfving, G. (1952), "Optimum Allocation in Linear Regression Theory," *Ann. Math. Statist.,* **23**, 255–262.

Fieller, E. C. (1954), "Some Problems in Interval Estimation," *J. Roy. Statist. Soc., Ser. B*, **16**, 175–185.

Karlin, S. and Studden, W. J. (1966), "Optimal Experimental Designs," *Ann. Math. Statist.*, **37**, 783–815.

Kendall, M. G. and Stuart, A. (1973), *The Advanced Theory of Statistics*, (Vol. 2, 3rd ed.), New York: Hafner.

McDonald, G. C. (1981), "Confidence Intervals for Vehicle Emission Deterioration Factors," *Technometrics,* **23**, 239–242.

McDonald, G. C. (1986), "Some Statistical Design Aspects of Estimating Automotive Emission Deterioration Factors," *Statistical Decision Theory and Related Topics IV* (Ed. S.S. Gupta and J.O. Berger), Vol. IV, Springer-Verlag, New York, pp. 363–371.

Searle, S. R. (1971), *Linear Models,* New York: John Wiley.

Studden, W.J. and D.J. VanArman (1969), "Admissible designs for polynomial spline regression," *Ann. Math. Statist.*, **40**, 1557–1569.

Appendix A. We will show here that under suitable conditions the limiting distribution of $Z_n = \sqrt{n}\left(\frac{\hat{\mu}_1}{\hat{\mu}_2} - \frac{\mu_1}{\mu_2}\right)$ is normal with mean zero and variance

$$\frac{1}{\mu_2^4}(\mu_2, -\mu_1)V\begin{pmatrix} \mu_2 \\ -\mu_1 \end{pmatrix}$$

as indicated in (2.19). We assume that $\mu_2 \neq 0$ and that $n^{-1}(X'X) = M_n \rightarrow M_o$ and that M_o is positive definite. In this $\hat{\mu}_2 = \hat{\beta}'\mathbf{f}(x_2) \rightarrow \mu_2$ in probability and we will define $Z_n = 0$ if $\hat{\mu}_2 = 0$. It is well known that the least squares estimates $\hat{\beta}$ are such that $\sqrt{n}(\hat{\beta} - \beta) \rightarrow N(0, M_0^{-1})$. The random variable Z_n can be written as

$$\begin{aligned} Z_n &= \sqrt{n}\ \frac{(\mu_2\hat{\mu}_1 - \mu_1\hat{\mu}_2)}{\hat{\mu}_2\mu_2} \\ &= \sqrt{n}\ (\hat{\mu}_2\mu_2)^{-1}(\mu_2, -\mu_1)\begin{pmatrix} \mathbf{f}'(x_1) \\ \mathbf{f}'(x_2) \end{pmatrix}(\hat{\beta} - \beta) \end{aligned}$$

In this case Z_n has a limiting distribution which is normal with mean zero and variance

$$\mu_2^{-4}(\mu, -\mu_1)\begin{pmatrix} \mathbf{f}'(x_1) \\ \mathbf{f}'(x_2) \end{pmatrix} M_0(\mathbf{f}(x_1), \mathbf{f}(x_2)\begin{pmatrix} \mu_2 \\ -\mu_1 \end{pmatrix}$$

as required.

Appendix B – Proof of Lemma 3.2. It suffices to minimize $R = (\mu_b, \mu_a)M_\ell^{-1}\left(\begin{smallmatrix} \mu_b \\ \mu_a \end{smallmatrix}\right)$ with respect to the information matrix M_ℓ restricted to be of the form (3.6). The subscript ℓ denotes that we are using the basis functions from equation (3.2). It is well known that we can restrict attention to designs on the end points a and b so that $M_\ell = \gamma_1 M_1 + \gamma_2 M_2$ where M_2 is diagonal with elements p_a and $1 - p_a$. It can readily be shown that the derivative of R with respect to p_a is

$$(-\mu_b, \mu_a)M_\ell^{-1}\begin{pmatrix} \gamma_2 & 0 \\ 0 & -\gamma_2 \end{pmatrix} M_\ell^{-1}\begin{pmatrix} -\mu_b \\ \mu_a \end{pmatrix}$$

Except for a positive factor of (det M_ℓ)^2 this is

$$\begin{aligned} &(\mu_b\gamma_1 m_{22} + \mu_b\gamma_2 - \mu_b\gamma_2 p_a + \mu_a\gamma_1 m_{21})^2 \\ &- (\mu_a\gamma_1 m_{11} + \mu_a\gamma_2 p_a + \mu_b\gamma_1 m_{12})^2 \end{aligned}$$

Since all the quantities are nonnegative an analysis of this factor shows the minimum of R to be given by p_a as expressed in Lemma 3.2 and equation (3.8).

Appendix C – Proof of Lemma 5.1 It suffices to minimize

$$(-\mu_1, \mu_1)\begin{pmatrix} \mathbf{f}'(x_1) \\ \mathbf{f}'(x_2) \end{pmatrix} M^{-1}(\mathbf{f}(x_1), \mathbf{f}(x_2))\begin{pmatrix} \mu_2 \\ -\mu_1 \end{pmatrix}$$

where $\mathbf{f}(x) = (1,\ x,\ (x-\xi)_+)$. As in the ordinary simple linear case we convert to a "Lagrange" basis similar to (3.2). The basis functions are based on a, ξ, b and are given by

$$\ell_a(x) = \begin{cases} \frac{\xi - x}{\xi - a} & x < \xi \\ 0 & x \geq \xi \end{cases}$$

$$\ell_\xi(x) = \begin{cases} \frac{x-a}{\xi - a} & x < \xi \\ \frac{b-x}{b-\xi} & x \geq \xi \end{cases}$$

$$\ell_b(x) = \begin{cases} 0 & x < \xi \\ \frac{x-\xi}{b-\xi} & x \geq \xi \end{cases}$$

Note that $\ell_a(x)$ has value one at $x = a$ and is zero at $x = \xi$ and b. Similarly for $\ell_\xi(x)$ and $\ell_b(x)$. It is known, see Studden and Van Arman (1969), that the optimal design must concentrate at a, ξ and b. In this case M_ℓ^{-1} is diagonal with diagonal elements p_a, p_ξ, p_b. In this case the above expression to be minimized reduces to $\gamma_a^2 p_a + \gamma_\xi^2 p_\xi + \gamma_b^2 p_b$ where $(\gamma_a, \gamma_\xi, \gamma_b)' = (\mathbf{f}(x_1), \mathbf{f}(x_2))\binom{\mu_2}{-\mu_1}$. Using Schwarz inequality the minimum is given by p_a, p_ξ, p_b proportional to $|\gamma_a|$, $|\gamma_\xi|$ and $|\gamma_b|$. Using the fact that $\mu_i = \mu_a \ell_a(x_i) + \mu_\xi \ell_\xi(x_i) + \mu_b \ell_b(x_i)$, for $i = 1, 2$; a small amount of algebra will show that these three numbers are proportional to w_1, w_2, w_3 given in (5.2).