FALCON RESEARCH AND
DEVELOPMENT

Falcon Research & Development Co.
A Subsidiary of Whittaker Corporation
One American Drive
Buffalo, New York 14225
716/632-4932

Whittaker

MASTER

FUEL ECONOMY OF AUTOMOTIVE VEHICLES
A COMPENDIUM OF STATISTICAL INTERPRETATIONS

October 1979

Report No. 8431/BUF-29

Prepared for

Department of Energy
U.S. Department of Energy
Conservation and Solar Energy
Transportation Programs

Under

Contract No. DE-AC-01-79CS-50084

DISCLAIMER

This book was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

Prepared by: H. T. McAdams

Approved by: Arthur Stein

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency Thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.

TABLE OF CONTENTS..

<u>Section</u>	<u>Title</u>	<u>Page</u>
1	INTRODUCTION	1
2	SUMMARY	3
3	DISCUSSION	5
	3.1 The General Motors Constant-Offset Thesis	7
	3.2 The Ford Interpretation	19
4	INTERPRETATION PAPERS	23
	Interpretation Paper No. 1--Implications of Change of Variables in the Regression Analysis of Vehicle Fuel Usage	24
	Interpretation Paper No. 2--Weighting-Factor Considerations in the Regression Analysis of Vehicle Fuel Usage	59
	Interpretation Paper No. 3--Weighting Factors Revisited: How They are Influenced by Assumptions Made in the Regression Model	91

1. INTRODUCTION

This report is submitted as the deliverable item in fulfillment of work performed under Contract No. DE-AC-01-79CS-50084 for the U.S. Department of Energy (DOE) in support of their evaluation of in-use and EPA fuel-economy data on automotive vehicles. The objective of the effort was to formulate statistical approaches to the data analysis and to assist in the interpretation of findings according to their impact on the Federal Fuel Economy Program (information and standards) and the overall energy conservation efforts of the DOE. The work was performed in coordination with another DOE contractor, Energy and Environmental Analysis, Inc., Arlington, VA.

The fuel economy of automotive vehicles as experienced by vehicle users is subject to a large number of variables. Driving profiles differ from one user to another, the mechanical condition of vehicles of the same make and model differ significantly, and a host of other circumstances conspire to cloud the fuel-economy picture as revealed by standard, EPA tests. To isolate and quantify the effects of all these variables on in-use fuel economy would require an extensive, well-structured sampling plan and copious data. Inasmuch as such sampling is usually impractical, it becomes necessary to draw from available data sources such inferences as can be supported by a combined application of statistical analysis and sound engineering judgment. As a result, many of the standard statistical techniques which would normally apply to a controlled and well-structured data base must be re-examined, and often reformulated, to meet the demands of data sources which fall short of what one would desire in a designed experiment. It was this function which was provided by the work performed under this contract.

Contractual requirements of the study, as set forth generically in the Statement of Work, were as follows:

1. Assist in the formulation of a statistical approach for analyzing year-to-year, test-to-test, model-to-model, test-to-in-use, car-to-truck, manufacturer-to-manufacturer, and technology-related trends in fuel economy.
 - a) Examine data sources and, where feasible, provide guidance in the formulation of sampling plans and/or experiment designs for data acquisition.
 - b) Where optimum statistical methods are not available, develop appropriate statistical techniques for the required analysis.

2. Critically examine conceptual bases for reporting and analyzing fuel-economy data with special attention to the use to which the results will be put.
3. Examine error propagation and the form of applicable error distributions, formulate appropriate statistical hypotheses, and develop and/or recommend appropriate tests of these hypotheses.
4. Assist in the application and interpretation of such analytical techniques as:
 - simple and multiple regression
 - weighted least-squares analysis
 - analysis of variance and covariance
 - regression analysis with both variables subject to error
 - rank order analysis by regression techniques.
5. Provide assistance and guidance in the presentation and explication of results according to the end use to which these findings are addressed.
6. Prepare working papers to document all phases of technique development, application and interpretation, and incorporate these working papers into an integrated final report.

This report summarizes the effort performed under the contract and is offered as a compendium of methodological approaches and viewpoints applicable to the assessment of in-use fuel economy of automotive vehicles.

2. SUMMARY

This report consists, for the most part, of three papers developed in chronological order. The term "Interpretation Papers" applied to these documents is particularly germane, inasmuch as the papers are aimed primarily at resolving differences of opinion which can arise from different points of view brought to the data. How such differences of viewpoint can arise is the subject of Section 3, Discussion.

Factors affecting the interpretation of data from fuel-economy studies include: (1) the choice of space in which to perform the analysis--e.g., miles per gallon, gallons per mile or some other alternative; (2) relative weights to be given to various items of data; and (3) characteristics of the error distribution assumed in the analysis. These factors and related issues are discussed in the three Interpretation Papers of Section 4.

Interpretation Paper No. 1 deals with "Implications of Change of Variables in the Regression Analysis of Vehicle Fuel Usage." Though miles per gallon has been used conventionally as a "figure of merit" for fuel economy, it is appropriate to consider as alternatives certain changes of variable, such as reciprocal miles per gallon (i.e., gallons per mile), or the logarithm of miles per gallon. Though various transformations can give essentially the same results within the interpolation range of the data, these transformations give divergent results when the corresponding regression equations are extrapolated beyond the range of the data. A further difficulty is involved in that each transformation implies a different weighting of errors, and this fact influences the least-squares error-minimization process conventionally used in developing the regression equation.

Interpretation Paper No. 2, "Weighting-Factor Considerations in the Regression Analysis of Vehicle Fuel Usage," examines factors affecting the precision and accuracy with which in-use fuel economy (fuel consumption) can be estimated by regression analysis. It is concluded that the assessment of fuel-economy shortfall for specific vehicle species imposes different statistical requirements than does the assessment of sales-weighted, average shortfall. Under the assumptions that errors are random with zero mean and constant variance, and that a linear regression equation (in the appropriate space) is valid, the paper develops criteria for best estimation of (1) the sales-weighted shortfall, and (2) shortfall as a function of EPA fuel economy.

Interpretation Paper No. 3 is entitled "Weighting Factors Revisited: How They Are Influenced By Assumptions Made in the Regression Model." The paper extends the considerations of the previous paper to the case in which the random errors have a finite expected value peculiar to each vehicle species.

Though these three papers by no means exhaust the questions which can arise in the interpretation of in-use, fuel-economy data, they provide information important to decision making in the subject area. Collectively, they address all of the salient points of the Statement of Work.

SAE Paper No. 790932, "Comparison of EPA and In-Use Fuel Economy of 1974-1978 Automobiles," exemplifies the application of the methodology of this report to actual fuel-economy data. The paper was a joint effort of DOE, EEA, and Falcon.*

* Paper by Barry McNutt (DOE); Robert Dulla (EEA) and H. T. McAdams (Falcon) presented at SAE Fuels and Lubricants Meeting, October 2-4, 1979, Houston, TX.

3. DISCUSSION

Standard tests to evaluate fuel economy or "gas mileage" can yield values which differ appreciably from values actually obtained from in-use vehicles, mainly because many factors which are controlled in the tests are essentially uncontrolled in the field. Driving profiles differ from one user to another and may depart significantly from the profile on which the standard test was based. In addition, the mechanical condition of a vehicle, whether it has or has not been adequately maintained, and a variety of other factors affect the extent of agreement between EPA test values for fuel economy and corresponding values derived from in-use experience. The comparison is further complicated by the fact that the discrepancy between test and in-use values can be quantified in various ways, depending on the purpose for which the comparison is made and the assumptions on which the comparison is based. As a result, different conclusions can be advanced from the same data, and it is important to be able to identify the source of these differences, even though it is recognized that certain affective bases for preferring one interpretation to another may remain. Whereas an optimist may observe that his glass is half full, the pessimist may still think of his glass as half empty. Similarly discordant views persist in the perception of the seriousness of the fuel-economy "shortfall."

The nature of such interpretational difficulties is well illustrated by the positions taken by General Motors, Ford, and DOE relative to questions posed by the House Subcommittee on Energy and Power of the Committee on Interstate and Foreign Commerce.¹ These questions arose from a DOE contractor report dated February 17, 1979 and entitled "On-Road Fuel Economy Trends and Impacts."² Major theses of this report were that: (1) a large discrepancy exists between on-road fuel economy

¹ Letter of March 1, 1979 from Congressman John D. Dingell, Chairman, Subcommittee on Energy and Power of the Committee on Interstate and Foreign Commerce, to Mr. John F. O'Leary, Deputy Secretary, U.S. Department of Energy.

² Energy and Environmental Analysis, Inc., "On-Road Fuel Economy Trends and Impacts," Final Report (February 17, 1979).

and the EPA certification composite value reported in the Gas Mileage Guide; (2) the discrepancy is greater for higher mpg cars; and (3) the discrepancy has grown appreciably from model year 1974 to model year 1977.

In their response to the Subcommittee, DOE reaffirmed its contention of a substantial shortfall and that "this gap is greater for higher mpg cars."³ General Motors, however, contended that "there appears to be a constant difference between fuel consumption measured on the EPA test versus that measured on the road" and "the situation is not getting worse from year to year."⁴ Ford offered data which, they asserted, "showed only a slight variation in the percentage relationship between EPA and in-use data for cars with differing EPA ratings" and expressed doubt that the shortfall was increasing with time.⁵ At the request of DOE, these differences in interpretation were examined in some detail in the interest of identifying basic sources of disagreement between the view held by DOE and the views held by GM and Ford.

³ Statement of John F. O'Leary, Deputy Secretary, U.S. Department of Energy, Before the Subcommittee on Energy and Power of the Committee on Interstate and Foreign Commerce (March 13, 1979).

⁴ General Motors Corporation Answers to Questions by the House Subcommittee on Energy and Power of the Committee on Interstate and Foreign Commerce (March 14, 1979).

⁵ Ford Motor Company Response to Congressman Dingell's Questions of March 5, 1979 Addressed to Mr. F. G. Secrest in Conjunction with Oversight Hearings to be Held March 13-14, 1979 (March 14, 1979).

3.1 The General Motors Constant-Offset Thesis

The thesis advanced by General Motors is that on-road performance should be characterized in terms of fuel consumption, expressed in gallons per mile, as opposed to fuel economy, expressed in miles per gallon. It is a further contention of GM that a constant offset of 0.01 gpm characterizes the relation between on-road and EPA-based fuel consumption. In support of the constant offset hypothesis GM develops regression equations in gpm space. These regression equations express on-road fuel consumption as linear functions of EPA fuel consumption.

If the usual assumptions with regard to regression analysis are made, the form of the regression equation would be

$$y = b_0 + b_1 x \quad (1)$$

where

x = EPA-based fuel consumption, gpm;
 y = on-road fuel consumption, gpm; and

b_0 and b_1 are constants to be determined. These constants are usually developed by applying the principle of least squares. In short, from among all possible values of b_0 and b_1 , one selects those particular values which minimize the quantity

$$E = \sum_{i=1}^N e_i^2 = \sum_{i=1}^N (y_i - b_0 - b_1 x_i)^2 \quad (2)$$

where x_i and y_i denote corresponding data values of EPA and on-road fuel consumption.

To minimize (2) it is customary to set the partial derivatives $\partial E / \partial b_0$ and $\partial E / \partial b_1$ equal to zero and solve the resulting normal equations for b_0 and b_1 . Note, however, that one might assign a fixed value for the intercept b_0 and then seek that value of the slope b_1 which minimizes the error function subject to the b_0 constraint. Alternatively, one might assign a value of b_1 to the slope and seek that value of b_0 which minimizes the error function

subject to the b_1 constraint. GM constrains b_0 to a value of 0.01 and finds that the least-square solution for b_1 generally is close to unity. Thus constraining the intercept to 0.01 is equivalent to constraining the slope to unity.

Let us consider the b_0 constraint and how it affects the solution for b_1 . If one differentiates (2) with regard to b_1 , one obtains

$$\begin{aligned}\frac{dE}{db_1} &= -2 \sum (y_i - b_0 - b_1 x_i) x_i = 0 \\ b_1 \sum x_i^2 - \sum x_i (y_i - b_0) &= 0 \\ b_1 &= \frac{\sum x_i (y_i - b_0)}{\sum x_i^2} \quad (3)\end{aligned}$$

Note that (3) does not provide a global minimum sum of squares of residuals but does provide the least sum of squares achievable for the assigned value of b_0 .

What is of especial interest in relation to the GM analysis is how b_1 changes in response to a change in b_0 . This question can be examined by differentiating (3) with respect to b_0 .

$$\frac{db_1}{db_0} = - \frac{\sum x_i}{\sum x_i^2} \quad (4)$$

In finite difference terms, (4) becomes

$$\Delta b_1 \approx - \frac{\sum x_i}{\sum x_i^2} \Delta b_0 \quad (5)$$

In other words, by adjusting b_0 downward by an increment Δb_0 , one can adjust b_1 upward by an increment Δb_1 . The proportionality constant is the quantity $\sum x_i / \sum x_i^2$.

Let us direct our attention to this proportionality constant. In consumption space the quantities x_i are much smaller than 1.0, a typical value being--say--0.05 gpm (corresponding to 20 mpg). The ratio

of x_i/x_i^2 for this typical value is 20. Note, also, that the ratio could range from 10 (for gpm = 0.1 or mpg = 10) to 40 (for gpm = 0.025 or mpg = 40). Typically the ratio of $\Sigma x_i/\Sigma x_i^2$ could be expected to be in the neighborhood of 15-25, therefore, in numerical value. In short, adjusting the magnitude of b_0 from--say--0.02 to 0.01 could increase the slope by as much as 0.25 (that is, from 0.75 to 1.00). Ample evidence of this fact is found in the GM equations (see attached Table 1). It suffices to cite only the overall equation

$$y = 0.02 + 0.83 x$$

which is adjusted to

$$y = 0.01 + 1.015 x$$

Note that for a change of 0.01 in the intercept the slope changes by 0.185, a ratio of 18.5! Unless there is some compelling reason to impose a constraint of 0.01 intercept, therefore, the adjustment must be regarded with concern, in view of its high-leverage effect on the slope. Note that unit slope is a key and crucial point in GM's argument that "real reductions in energy consumption are being realized on the road and these reductions are happening at a rate equal to that measured on the EPA test."

Let us examine the implications of equation (5) in mpg space. Now the values of the x_i are much greater than 1.0, a typical value being 20. For this value, the ratio x/x_i becomes 0.05. The range to be expected is from 0.10 (for 10 mpg) to 0.025 (for 40 mpg). In other words, in order to drive the slope of the fuel-economy regression line from 0.75 to 1.00, it would be necessary to drive the intercept downward by perhaps 20 times 0.25, or as much as 5 mpg. Such adjustment would most likely produce a negative intercept! It is clear, therefore, that the imposing of constraints on the regression equation produces very different effects in gpm space than in mpg space and that the use of such constraints can, under an appropriate choice of space, strongly affect conclusions regarding whether or not shortfall is increasing with time.

Table 1: GM REGRESSION EQUATIONS

$$y = a + bx$$

Model Year	N	<u>Unconstrained Equation</u>			<u>Constrained Equation</u>		
		<u>a</u>	<u>b</u>	<u>r</u>	<u>a</u>	<u>b</u>	<u>r</u>
<u>Post Card Survey</u>							
75	13	0.004	1.104	0.96	0.01	1.013	0.95
76	13	0.013	0.980	0.93	0.01	1.038	0.92
78	47	0.015	0.910	0.94	0.01	1.010	0.93
<u>Peterson, Howell & Heather</u>							
75	28	0.022	0.790	0.85	0.01	0.968	0.82
76	35	0.013	0.950	0.91	0.01	1.003	0.91
77	34	0.022	0.800	0.89	0.01	1.002	0.85
78	33	0.024	0.794	0.81	0.01	1.060	0.79
79	28	0.023	0.819	0.79	0.01	1.072	0.73
<u>All Data</u>							
	231	0.020	0.830	0.90	0.01	1.015	0.88

An often-used measure of the "goodness of fit" of a regression equation is the correlation coefficient r , or its squared value r^2 . The latter quantity expresses that fraction of the total sum of squares which is "explained by" the regression equation, and $1-r^2$ expresses that fraction of the total which is associated with "residuals"--that is, departures from the regression line. The "total" sum of squares is computed with reference to the mean value of the dependent variable.

Note that the 0.01 intercept constraint causes only a slight decrease in the value of r . For that reason, one is tempted to argue that the constrained equation can be considered to be as "reasonable" as the unconstrained equation. By virtue of the fact that the two equations tend to cross near the center of the cloud of points, however, the residuals referenced to the constrained line are not much different from the residuals referenced to the unconstrained line. The differences which do occur tend to be obscured by the fact that the total sum of squares is of much greater magnitude than either of the residual sums of squares, and it is this fact that makes r relatively insensitive to the constraint. The implications of the change in equations, however, are very important and should not be decided on the basis of r , which at best is a poor measure of the "relation" between two variables.

The insensitivity of r to adjustments in the values of intercept and slope can be examined by a straightforward argument. Let $\bar{y}$ be the average of all the on-road numbers. Then the total sum of squares is

$$\text{Total SS} = \sum_{i=1}^N (y_i - \bar{y})^2 \quad (6)$$

and, by (2), the residual sum of squares, for any choice of b_0 and b_1 , is

$$\text{Residual SS} = \sum_{i=1}^N (y_i - b_0 - b_1 x_i)^2 \quad (7)$$

Finally, the regression sum of squares is

$$\text{Regression SS} = \text{Total SS} - \text{Residual SS}$$

$$= \sum_{i=1}^N (y_i - \bar{y})^2 - \sum_{i=1}^N (y_i - b_0 - b_1 x_{i0})^2 \quad (8)$$

and the squared correlation coefficient is

$$r^2 = \frac{\sum_{i=1}^N (y_i - \bar{y})^2 - \sum_{i=1}^N (y_i - b_0 - b_1 x_i)^2}{\sum_{i=1}^N (y_i - \bar{y})^2} \quad (9)$$

$$= 1 - \frac{\sum_{i=1}^N (y_i - b_0 - b_1 x_i)^2}{\sum_{i=1}^N (y_i - \bar{y})^2}$$

How r^2 changes in response to changes in b_0 and b_1 can be exhibited by using (9) to generate "equal- r " contours in the (b_0, b_1) -plane. The maximum value of r^2 is achieved at only one point in the plane, that point at which b_0 and b_1 satisfy the least-squares condition. As one decreases r^2 , however, the envelope of allowable values of b_0 and b_1 becomes larger. If the r^2 "surface" is relatively flat in the vicinity of the least-squares values of b_0 and b_1 , then adjustment of these constants to sub-optimum values induces only a small decrease in r^2 .

To investigate this possibility one can synthesize data having essentially the same properties as the General Motors data and observe how r changes in response to changes in the constrained value of the intercept. The data shown in Figure 1 are offered for this purpose.

FIGURE 1 Hypothetical Regression Relation Between EPA and On-Road Fuel Consumption

The effect of constraining the regression line to fixed intercepts is shown in Figure 2. Note that intercepts of 0.01 and 0.03 do not appreciably degrade the correlation coefficient. Further explication of this point is made in Figure 3, which shows that, in the vicinity of the "best fit" value of 0.02 for the intercept, the correlation coefficient is relatively insensitive to perturbation about that value. In particular, an intercept of 0.03 would seem to be as logical as an intercept of 0.01 if one were to use the correlation coefficient as the basis of choice. Accordingly, one must look elsewhere for a basis for choice of intercept.

For this purpose it is appropriate to consider a confidence-bound approach. Figure 4 provides a 3-sigma confidence band for the regression equation. The band is narrowest at the mean value of the EPA fuel consumption for the data, 0.052 gpm, and diverges as one moves toward either higher or lower values. At the origin, the confidence bounds are those applicable to the intercept; consequently, on the assumption of normally distributed error, it can be asserted that 0.02 ± 0.004 is a 99.7% confidence bound for the intercept. For the hypothetical data, therefore, there would seem to be little statistical justification for constraining the intercept to values outside the indicated bounds.

Whether one chooses to accept GM's constrained regression equations has serious implications with regard to the question of whether shortfall is increasing with time. Compare, for example, the previously cited unconstrained equations with the equation with 0.01 intercept constraint:

$$y = 0.02 + 0.83 x \quad (\text{unconstrained})$$

$$y = 0.01 + 1.015 x \quad (\text{constrained})$$

In Table 2 the gpm and corresponding mpg values are computed by the two equations and the shortfalls computed in both spaces. Finally, the shortfalls are expressed as percents of the corresponding EPA-based gpm and mpg respectively. If shortfall is interpreted as the increased fuel required to travel a mile of distance, it is clear that unit slope is essential to the argument that shortfall does not depend on EPA-based fuel economy. If the slope is less than one, however, the shortfall increases with EPA miles per gallon. Conversely, if the slope is greater than one, the shortfall actually decreases with increasing mpg. To the extent that the average EPA-based mpg is expected to increase with time, unit slope assures that there will be no increase in shortfall due to EPA-based CAFE.

FIGURE 2 Alternative Choices of Regression Equation

FIGURE 3 Effect of Constrained Intercept on Slope and Correlation Coefficient

FIGURE 4 Confidence Interval for On-Road Fuel Economy as Computed by Regression

This consideration is separate from the question of technology trends, however. Even if unit slope is maintained over time, there is no reason that the intercept (the postulated "constant offset") could not increase or decrease with time. Constraining the intercept to 0.01 for the several model years (see Table 1) essentially imposes this condition and, by virtue of the effect of this constraint on the slope, essentially imposes unit slope also! Note, however, that even if these adjustments are made, the fact remains that percent shortfall will indeed increase with time, if for no other reason than the upward trend of CAFE, whether the percent is computed in gpm or mpg space and whether it is computed with constrained or unconstrained equations.

In summary, the imposing of constraints on the regression analysis of fuel-economy data can have a high leverage effect on conclusions. Unless the 0.01 intercept constraint can be defended on the basis of engineering arguments or on the basis of a statistical hypothesis test, it must be viewed with caution. Similar considerations apply in deciding whether the regression analysis should be performed in mpg or gpm space and the extent to which data from various sources are weighted in the analysis. These issues are addressed in the Interpretation Papers which constitute the bulk of this report.

3.2 The Ford Interpretation

In the preceding section of this report, the approach to data analysis taken by GM was cited to highlight the extent to which choice of space (whether mpg or gpm) and the use of constraints can affect conclusions drawn from the data. Further light is shed on these issues by an examination of the Ford response to the Subcommittee questions (op. cit.) and a previously published paper by Ford on in-use fuel economy.⁶

It will be recalled that GM argues strongly for analyzing vehicle fuel-use data in terms of gallons-per-mile (gpm) rather than miles-per-gallon (mpg). By working in fuel-consumption space and advancing the hypothesis of a constant gpm "offset," GM was able to assert that the objectives of the fuel-economy improvement program are being met on schedule. Eschewing this argument, Ford prefers to perform their analyses in mpg space and computes shortfall as a percentage of the EPA-rated fuel economy. Both the GM and Ford approaches tend to modify the perceived seriousness of the fuel-economy shortfall but in different ways.

Consider for example, the fact that the difference between EPA and in-use fuel economy tends to increase with increasing EPA rating. GM's assertion of a constant "offset" in gpm space appears to erase this tendency. Note, however, that an 0.01-gpm offset amounts to a considerably larger mpg shortfall for a vehicle rated at 30 mpg than for a vehicle rated at 20 mpg. The process of taking reciprocals of the mpg numbers thus is essentially a "stabilizing transformation" in the sense that it tends to make the shortfall for all vehicles "look the same."

The percentage transformation has a similar stabilizing influence. For example, a 3-mpg shortfall at 15 mpg rating and a 6-mpg shortfall at 30 mpg rating both map into a 20 percent shortfall, although in magnitude one is twice as great as the other. Both of these transformations, or any others which tend to have such a "leveling" effect, should be used with caution, because the "leveling" tends to affect all conclusions pertaining to whether the shortfall is a constant or non-constant quantity. For example, if a transformation tends to minimize the increase of shortfall with EPA rating, it may also tend to minimize any increase of shortfall with time, if such an increase exists. In this light, one should view with caution the following conclusion of South and Raja (op. cit.).

"The percentage difference between in-use and EPA mpg rating is relatively constant with the level of mpg. Also, this difference is relatively constant for different classes of cars."

⁶ N. E. South and R. Raja, "In-Service Fuel Economy," SAE Paper No. 790227 (1979).

Though it may be entirely appropriate to view shortfall as an effect proportional to mpg, it is easy mentally to identify "constant percentage shortfall" with "constant shortfall."

With regard to weightings to be applied to data, it is of interest to examine Ford's criticism of the DOE analysis as well as their own approach to the weighting question. For example, in their response to the Subcommittee (op. cit.), it is stated that the DOE analysis "is subject to technical correction" in that "all points were used with equal weighting, although one data point may have represented hundreds of cars and another only one or two cars." That this criticism has some validity is made clear in Interpretation Paper No. 2, but the further considerations of Interpretation Paper No. 3 make evident the fact that there is no easy answer to the choice of weighting factors. Some of the difficulties which can be encountered can be illustrated by the Ford analysis.

Consider, for example, the procedure employed by South and Raja (op. cit.) in their culling of samples to be included in their data analysis. No model type was considered if its sample size was less than 10 vehicles, but sample size varied widely (up to more than 400 vehicles for some model types). If the averages for each model type are used as data inputs for regression analysis, these averages have widely different inherent variances, and some form of weighted least-squares analysis would be required to take this variation into account. The paper fails to clarify this point, as well as the basis for computing the residual (error) variance for the regression line which, as noted, would be affected by the variable sample sizes employed in the analysis. The confidence bounds given in the paper, therefore, need clarification.

Weighting difficulties are further suggested by the method employed by South and Raja in their averaging of summer and winter results. Observing that the miles driven per day is 58 in summer and 38 in winter, they compute a 60/40 weighted harmonic average. First, let it be noted that average miles per gallon is the ratio of total miles driven to total gallons of fuel consumed. A more appropriate partitioning, therefore, would be one based on the total miles driven in summer as compared to total miles driven in winter, and these values can not be obtained from per diem values alone. The missing factor is the length of summer (in days) as compared to the length of winter.

This point can be emphasized by hypothesizing two situations, one in which winter constitutes 30% of the year, the other in which winter constitutes 70% of the year (see Table 2). For the short-winter year, only 22% of all miles are accumulated under winter conditions, not the 40% which characterizes a typical winter day. For the long-winter year,

61% of all miles are accumulated under winter conditions, not the 40% postulated on the basis of a typical day. For the per diem 60/40 split to be valid on an annual basis, it must be postulated that winter and summer are of equal length. A second criticism pertains to the use of harmonic averaging in this instance. It will be recalled that a harmonic average is appropriate when equal number of miles are represented by the classes being averaged. This assumption is explicitly rejected here by the assertion of 60 miles per day in summer and 40 miles per day in winter, unless summer constitutes exactly 40% of the days and winter 60%.

Table 2 CALCULATION OF SUMMER/WINTER WEIGHTING FACTORS

	Short Winter (30% of all days)		Long Winter (70% of all days)	
		<u>Percent of Miles</u>		
Winter	$0.3 \times 0.4 = 0.12$	(22%)	$0.7 \times 0.4 = 0.28$	(61%)
Summer	$0.7 \times 0.6 = \underline{0.42}$	(78%)	$0.3 \times 0.6 = \underline{0.18}$	(39%)
	.54		.46	

How, then, should the summer and winter results be combined? An example will suffice to answer this question. Three items of information are necessary: miles per gallon (or, gallons per mile), both summer and winter; lengths of seasons (days); number of miles driven per day, summer and winter. Assume:

winter 30% of days; summer 70%

winter 40 miles per day; summer 60 miles per day

winter 15 mpg; summer 18 mpg.

Calculate:

Total winter miles = $0.3 \times 365 \times 40 = 4380$

Total summer miles = $0.7 \times 365 \times 60 = \underline{15330}$

Total miles per year 19710

Total winter gallons = $4380 \div 15 = 292$

Total summer gallons = $15330 \div 18 = \underline{852}$

Total gallons per year 1144

Annual average miles per gallon = $19710 \div 1144 = 17.2$

Compare this figure with 16.7 mpg computed as a 60/40 harmonic average or 16.5 as a simple (equally weighted) arithmetic average. It might be concluded that the method of averaging is not crucial, but the details are presented to point up the nature of inconsistencies which can arise in the use of weighting factors. A final remark in this connection can be made with regard to how South and Raja "average" the regression lines. Again they average harmonically by selecting two points each on the summer and winter curves. This does not necessarily assure a least-squares solution and gives a solution which depends to some extent on how the points were selected.

In summary, the use of transformations and weighting factors in data analysis can influence conclusions drawn from the data. For that reason, it is important to have a well-thought-out protocol for such matters and to be cognizant of the assumptions on which that protocol is based. The three Interpretation Papers which constitute the bulk of this report are aimed at providing background theory and information important to the making of these procedural decisions.

4. INTERPRETATION PAPERS

In this section the three Interpretation Papers evolving from this study are presented in their chronological order of development. No attempt is made to unify the notation in these papers, since each paper stands as an independent document within which notation is internally consistent. The same comment applies to the fact that appendices, where they occur, are attached to the working papers individually rather than to the report as a whole. Figures, tables and lists of references also apply to individual working papers, no attempt having been made to coordinate these items throughout the report as a whole.

Whittaker

Interpretation Paper No. 1
Project No. 8431
US Department of Energy
June 8, 1979
H. T. McAdams

IMPLICATIONS OF CHANGE OF VARIABLES IN THE REGRESSION
ANALYSIS OF VEHICLE FUEL USAGE

1. Introduction

Miles per gallon and gallons per mile are two quantities used to characterize the intensity of fuel usage by automotive vehicles, but these quantities by no means exhaust all the possible ways in which fuel usage can be viewed. Vehicle speed is an important factor, and a vehicle consumes fuel even when idling and no miles are being accumulated. Why not, then, gallons per hour or hours per gallon? Moreover, heavy vehicles generally use more fuel in a given scenario than do light vehicles, but the performance of the two may be essentially the same if considered on the basis of "pound-miles per gallon." Finally, even such subjective considerations as "quality of ride" and "aesthetic appeal" conspire to differentiate a "Cadillac-mile" from a "Rabbit-mile" and suggest that distance traveled is not the only value purchased in a gallon of gasoline. How to select the most appropriate figure of merit for fuel-usage efficiency, therefore, is by no means obvious, even though miles per gallon and, somewhat more recently, gallons per mile have evolved as "natural" choices.

Because of the emphasis on resource conservation, it has become vogue to speak of "fuel economy" rather than "fuel consumption," a practice decried in a recent editorial in Automotive Engineering.⁽¹⁾

(1) Larry Givens, "Conspicuous Economy?", editorial in Automotive Engineering, Volume 87, Number 1 (January 1979).

"The word economy implies a reduction in consumption, a lessening of a previous rate of consumption. Thrift, frugality, even parsimony can be considered synonyms for economy. Consumption, on the other hand, means depletion or exhaustion—a simple using-up of something. The DOE's Gas Mileage Guide lists fuel *consumption* of automobiles in miles per gallon; however, comparing this year's Guide with those of previous years can show the improved fuel *economy* of 1979 cars, since the fuel consumption of a 1979 model is less than that of its earlier-year equivalent in many cases."

Further obfuscation has resulted from a present tendency to use the term "fuel economy" when speaking of miles per gallon and "fuel consumption" when speaking of gallons per mile. It should be clear, however, that both are measures of the same intrinsic property--namely, the intensity of fuel usage by an automotive vehicle. Both quantities can be said to be "metameters" for expressing that property in quantitative terms, since one can be readily transformed into the other.

In spite of the above logic, it does make a difference whether statistical analysis of fuel usage is performed in so-called fuel-economy space (miles per gallon) or in fuel-consumption space (gallons per mile). This difference is evident when one compares arithmetic and harmonic averages for miles per gallon for a collection of vehicles.⁽²⁾ The lack of agreement between the two averages arises because the correspondence between miles per gallon and gallons per mile for individual vehicles is not preserved when the quantities involved are subjected to additional computation. In mathematical terms, the collection of miles-per-gallon figures and the collection of gallons-per-mile figures do not constitute an isomorphism. For two collections to be isomorphic their elements must not only be in one-to-one correspondence but that correspondence must be preserved under appropriately defined mathematical operations. Such is not the case for miles-per-gallon and gallons-per-mile even in such a simple operation as averaging, and the discrepancy becomes even more vexing when in-use fuel-consumption (fuel economy) values are regressed on the corresponding EPA values.

It is at this point that the statistical analyst faces a dilemma. On the one hand it is admitted that miles per gallon and gallons per mile are only "aliases" for the same property of a vehicle, the efficiency with which it converts fuel into transportation. On the

(2) H. T. McAdams, Fuel Economy of Automotive Vehicles--A Compendium of Statistical Techniques, Falcon R&D Co., November 1977 (Working Paper No. 2)

other hand, analyses performed in the two spaces do not give the same results and can lead to different conclusions. Which, then, is "correct" --or is either? One is left with the lurking suspicion of statistical bankruptcy, the feeling that the analysis, however performed, is lacking in substance if it can be so easily warped by the predilection of the analyst.

It is the purpose of this paper to examine this dilemma from both fundamental and practical points of view. Attention will be given not only to the miles-per-gallon/gallons-per-mile controversy but to alternative metameters and figures of merit for characterizing vehicle fuel efficiency.

2. Mathematical and Statistical Aspects of Fuel-Use Metameters

Fuel-use data can be obtained by means of the EPA certification test or by other tests employing either a dynamometer rig or actual road experience. The so-called fuel-economy "shortfall" is concerned with the relation between in-use experience and EPA fuel-economy ratings.

Consider a set of automotive vehicles $A = \{a_1, a_2, \dots, a_n\}$. If one associates with the i th vehicle an EPA number x_i and an in-use number y_i , one obtains a set of ordered pairs (x_i, y_i) from which one can hopefully develop a function $y = f(x)$. This function would map the EPA value x_i , as preimage, into an in-use number y_i , as image.

In reality the relation between in-use and EPA numbers is a many-to-one mapping. For a given certification value x , an entire distribution of in-use values arises. Each vehicle belonging to a certification class of vehicles has its own peculiarities, and is most likely driven at different speeds, over different terrain, and under a host of other conditions which are different from those to which all other vehicles of the class are subjected. A reasonable approach is to regard these peculiarities as random variables hopelessly beyond control and to abstract from this array of many-to-one correspondences an idealized functional relation which minimizes, in some sense, the uncertainty inherent in the prediction of y , given a measured value of x . Typically this minimization is achieved by least squares, as employed in regression analysis.

The function $y = f(x)$ can conceivably be of arbitrary form, but it has been customary to assume it to be linear in x . Let it be noted, however, that under a change of variables a linear relation does not,

in general, transform into a linear relation. It is for this reason that the various metameters capable of being employed to represent fuel-use intensity may give rise to different functional relationships. The constraint of linearity can, of course, be imposed on the data for a variety of choices of metameters but it would be fortuitous, to say the least, if the resulting functions were equivalent. Additional complication is introduced by the fact that any transformation of variables also has the effect of transforming the error distribution associated with the in-use values observed for a given EPA value.

The mathematical and statistical consequences of a change in variables will be considered in the following two subsections of this paper. In the mathematical discussion it will be assumed that the function $y = f(x)$ is error free and that, under a change of variables, the function changes form in strict accordance with mathematical law. The statistical discussion will explore the effect of the transformation on the estimation process.

2.1 Mathematical Consequences of a Change of Variables

Let it be assumed that a set of measurements have been employed to derive a linear regression equation expressing in-use fuel economy (miles per gallon) as a linear function of EPA certification fuel economy. Thus

$$y = a + bx \quad (1)$$

where a is the y-axis intercept* and b is the slope of the regression line. Consider the change of variables

$$\left. \begin{array}{l} Y = 1/y \longrightarrow y = 1/Y \\ X = 1/x \longrightarrow x = 1/X \end{array} \right\} \quad (2)$$

where, following current terminology, X and Y denote fuel consumption (gallons per mile). Substitution of (2) in (1) gives

$$1/Y = a + b (1/X)$$

* Note that the notion of a y-axis intercept is physically intractable because the function is undefined at the origin. The EPA test requires a finite number of miles traveled and, consequently, a finite number of miles per gallon. Though this value can be numerically small, it can not attain zero magnitude.

or, upon simplification,

$$Y = \frac{X}{b + aX} \quad (3)$$

Thus if miles per gallon obeys a linear relation, then certainly gallons per mile does not, unless $a = 0$.

Suppose, now, that one desires to invoke a linear relation between Y and X at the expense of some error in prediction. Clearly the equation

$$Z = c + dX \quad (4)$$

can be made to agree with (3) at two selected values of X . Denoting these two values X_1 and X_2 and equating corresponding values of Y and Z , one obtains

$$\left. \begin{aligned} c + d X_1 &= Y_1 \\ c + d X_2 &= Y_2 \end{aligned} \right\} \quad (5)$$

with the solution

$$\left. \begin{aligned} d &= \frac{Y_1 - Y_2}{X_1 - X_2} \\ c &= Y_1 - d X_1 \end{aligned} \right\} \quad (6)$$

As an example, suppose that in mpg space

$$y = 1.83 + 0.73x \quad (7)$$

Then, by virtue of (3),

$$Y = \frac{X}{0.73 + 1.83X} \quad (8)$$

in gpm space. It is desired to develop a linear form for (8) by imposing constraints at $x = 10$ ($X = 0.1$) and $x = 25$ ($X = 0.04$). Then $Y_1 = .1095$, $Y_2 = 0.0498$ and

$$d = \frac{0.1095 - 0.0498}{0.1 - 0.04} = \frac{0.0597}{0.06} = 0.995$$

$$c = 0.1095 - (0.995)(0.1) = 0.01$$

The solution is

$$Z = 0.01 + 0.995X \quad (9)$$

with

$$Z_1 = 0.01 + 0.995(0.1) = 0.1095$$

and

$$Z_2 = 0.01 + 0.995(0.04) = 0.0498$$

in agreement with Y_1 and Y_2 as postulated.

Note, also, that

$$\frac{1}{Y_1} = \frac{1}{Z_1} = \frac{1}{.1095} = 9.13 \text{ mpg}$$

and

$$\frac{1}{Y_2} = \frac{1}{Z_2} = \frac{1}{0.0498} = 20.08 \text{ mpg}$$

in agreement with

$$y = 1.83 + 0.73x.$$

However, if a different point--say $x = 17.5$ ($X = 0.0571$)--is selected,

and

$$y = 14.6$$

$$Y = \frac{1}{14.6} = 0.0685.$$

By equation (8), however,

$$Z = 0.01 + 0.995(0.0571) = 0.0668$$

and $1/Z = 15.0$ which departs from the postulated value of $y = 14.6$. The difference of 0.4 mpg is essentially the maximum difference which could occur between the limits of $x = 10$ and $x = 25$. Greater deviations could occur outside this range, however. For example, for $x = 35$ mpg ($X = 0.02857$ gpm), equation (7) predicts $y = 27.38$ mpg ($Y = 0.0365$ gpm). Equation (9) predicts $Y = 0.03843$ gpm ($y = 26.02$ mpg) with a deviation of 1.36 mpg.

Now, it is pointed out that the development of a linear equation in gallons-per-mile space would not, in general, be accomplished by constraining the equation at two points. Rather, it would be developed by least-squares error minimization in gpm space. This procedure, however, is only a different type of constraint and a constraint different from least-squares error minimization in mpg space. More about this matter will be said in a later section.

2.1.1 Is There a Constant Shortfall in gpm Space?

In the above example the slope of the assumed linear relation in miles-per-gallon space transformed into a near-unity slope in gallons-per-mile space. Suppose it should turn out that this near-unity slope tends to prevail for a variety of miles-per-gallon regression relations. Then it would appear that there is a constant difference between in-use fuel consumption and certification fuel consumption independent of size, weight and other determinants of vehicle fuel requirements. This hypothesis has been advanced by General Motors in their answer to questions by the House Subcommittee on Energy and Power.*

* General Motors Corporation Answers to Questions by the House Subcommittee on Energy and Power of the Committee on Interstate and Foreign Commerce on Fuel Economy Program Established by the Energy Policy and Conservation Act and Related Matters. Washington, DC. (March 14, 1979).

To consider the reasonableness of this hypothesis let us return to equation (6). Note that

$$d = \frac{y_1 - y_2}{x_1 - x_2} = \frac{\frac{1}{y_1} - \frac{1}{y_2}}{\frac{1}{x_1} - \frac{1}{x_2}} = \frac{y_2 - y_1}{x_2 - x_1} \cdot \frac{x_1 x_2}{y_1 y_2}$$

or

$$d = b \frac{x_1 x_2}{y_1 y_2} \quad (10)$$

Setting $d = 1$, one has

$$b \frac{x_1 x_2}{y_1 y_2} = 1$$

or

$$b = \frac{y_1}{x_1} \cdot \frac{y_2}{x_2} \quad (11)$$

Thus it is clear that for selected values of x_1 and x_2 it is possible to maintain values of $d = 1$. Note that (11) can be written

$$b = \frac{a + b x_1}{x_1} \cdot \frac{a + b x_2}{x_2} \quad (12)$$

Suppose one lets $x_1 = 10$, $x_2 = 25$. Then (12) becomes

$$b = \frac{a + 10b}{10} \cdot \frac{a + 25b}{25}$$

$$250b = a^2 + 35ab + 250b^2$$

$$250b^2 + (35a - 250)b + a^2 = 0 \quad (13)$$

Upon solution of (13) one obtains

$$b = 0.5 - 0.07a \pm 0.002 \sqrt{(35a - 250)^2 - 1000} \quad (14)$$

which provides a contour of combinations of slope and intercept in mpg space which map into unit slope in gpm space. For example, if $a = 1.8$,

$$\begin{aligned} b &= 0.5 - 0.07(1.8) \pm 0.002 \sqrt{[(35)(1.8) - 250]^2 - 1000} \\ &= 0.5 - 0.125 \pm 0.369 \\ &= 0.744, 0.006 \end{aligned}$$

Note that the value of 0.744 is essentially in agreement with a case previously cited ($a = 1.83$, $b = 0.73$, $d = 0.995$).

The conditions set forth above specify values of a and b which lead to unit slope in gallons-per-mile space. What is of more interest, however, are conditions which lead to a constant intercept in that space, since it is the General Motors contention that this intercept does not vary over time.

The general linear expression for in-use miles per gallon is

$$y = a + bx \quad (1)$$

and this leads to

$$Y = \frac{X}{b + aX} \quad (3)$$

in gallons-per-mile space. For the difference between in-use gallons-per-mile (Y) and certification gallons-per-mile (X) to be some constant--say k --it is necessary that

$$\frac{X}{b-aX} - X = k \quad (15)$$

From (15) one can constrain X to two values to obtain two equations which can be solved for a and b in terms of k . For the resulting ensemble of (a, b) -combinations, the constant difference would hold at the end-points of a range of X values ($X_1 \leq X \leq X_2$) and presumably would hold approximately at all interior points of the interval.

By substituting the two interval end-points X_1 and X_2 in (15) and solving for a and b (see Appendix I), one obtains

$$\left. \begin{aligned} a &= \frac{B_1 X_2 - B_2 X_1}{A_2 B_1 - A_1 B_2} \\ b &= \frac{A_2 X_1 - A_1 X_2}{A_2 B_1 - A_1 B_2} \end{aligned} \right\} \quad (16)$$

where

$$\left. \begin{aligned} A_1 &= kX_1 + X_1^2 \\ A_2 &= kX_2 + X_2^2 \\ B_1 &= X_1 + k \\ B_2 &= X_2 + k \end{aligned} \right\} \quad (17)$$

For example, if one takes $k = 0.01$ as postulated by General Motors, one obtains

$$a = 1.82$$

$$b = 0.73$$

with $X_1 = 0.1$ gpm ($x = 10$ mpg) and $X_2 = 0.04$ gpm ($x = 25$ mpg). At $x = 17.5$ mph, halfway between 10 and 25 mpg, the difference between in-use and EPA gallons per mile is 0.0114, as compared with 0.01 at the interval end points, 10 and 25 mpg. A second example, for $k = 0.015$, yields

$$a = 2.37$$

$$b = 0.63$$

for the same end-points of 10 mpg and 25 mpg, and other combinations can be easily explored.

In conclusion, it can be stated that there is a multiplicity of miles-per-gallon regression equations which can map into a constant gallons-per-mile shortfall. Whether it can be argued that this difference is constant over time must be approached with care, however. Because of their small absolute values, changes in gallons per mile are disarming. Deviations from a postulated shortfall of--say--0.01 gpm do not have to be of large magnitude to be of economic significance. Note that an increase from 0.01 to 0.015 is a 50% increase in the shortfall and, relative to a vehicle rated at 20 mpg, represents a drop from 16.7 mpg to 15.4 mpg. Even if it is argued that the constant difference of 0.01 prevails over time, it must be recognized that, on a relative basis, the effect of this difference becomes more pronounced as gallons-per-mile figures decrease toward the 1985 target of 0.0364 gpm corresponding to 27.5 mpg. The 0.01-gpm discrepancy represents a 27.5% overage of fuel consumption referenced to 27.5 mpg in comparison with only a 20% overage when referenced to 20 mpg.

2.1.2 If There is a Constant gpm Shortfall--Why?

Let it be assumed that a constant difference between in-use and EPA-rated fuel consumption exists, according to the relation

$$Y = 0.01 + X$$

where X and Y are expressed in gallons per mile. The implication of this assumption is that a luxury vehicle rated at 10 mpg and a subcompact vehicle rated at 35 mpg both have fuel-consumption overages of 0.01 gallons per mile. A rationale for this phenomenon, if real, would seem to require a component of fuel expenditure which does not depend substantially on the operating efficiencies of the vehicles. Note, however, that the 0.01 gpm overage is only 10% of the rated gpm for the luxury vehicle but is 35% of the rated gpm for the subcompact.

Several sources of fuel expenditure can be identified which are partially, if not totally unproductive of mileage accumulation. Fuel spilled or lost by evaporation produce no vehicle transport and are consequently candidate sources of unproductive fuel expenditure. Moreover, logic suggests that these sources would not be highly vehicle dependent. Time spent at idle also contributes no mileage accumulation.

Consequently, if the fraction of time spent in idle mode is substantially greater than that allotted in the EPA test, then an additional increment of "unproductive fuel" would be contributed to the fuel-consumption budget. Though a large engine would tend to consume more fuel per hour while idling than a small vehicle, this difference would undoubtedly be less than in the case of operation of the two vehicles for unit time at a speed near that for maximum efficiency. A similar situation would occur at "very low" and "very high" speeds, inasmuch as the curves relating fuel consumption and speed for various vehicles tend to converge at extreme speeds. Thus if the speed distribution of in-use vehicles tends to be bimodal and to depart substantially from the distribution assumed in the EPA test, then fuel consumed would tend to exceed that used in the standard test. To the extent that differences among vehicles are minimized at extreme speeds, this component of vehicle consumption might tend to be relatively uncritical of design differences among vehicles. Comfort accessories, such as heaters and air conditioners, could add fuel increments across the board, and the quality of the fuel itself could be a factor. If the average grade of fuel in the field is of lower BTU content than that used in the EPA test, a corresponding increase in fuel consumption would be experienced.

Finally, human-factor considerations may be more important than is generally realized. It is generally conceded that drivers differ appreciably in their "aggressiveness," and that these differences can be reflected in fuel lost in fuel-wasteful acceleration and deceleration maneuvers. Though it might be argued that these effects might "average out" over the user public, there is reason to believe that such would not be the case. In the EPA test the driver "follows the chart" and can--indeed, must--anticipate maneuvers required to keep schedule. In the real world a very different situation prevails. The driver "closes the control loop" via visual cues received from the traffic of which he is a part and can not anticipate required maneuvers in the same way as can the test driver. Moreover, the test subjects every vehicle to the same schedule. In comparable traffic there would be no passing, no lane-switching and no queuing. In short, there is variability in real-world driving which is not modeled in the test, and variability is postulated to have an adverse effect on fuel economy, via delays, accelerations and decelerations, and human-factor variables. Many of these considerations tend to be substantially vehicle independent.

In summary, it is possible to identify certain sources of fuel expenditure which are more or less independent of vehicle design. It does not necessarily follow, however, that these sources insure that a constant difference between in-use and EPA-rated gallons per mile will persist over time.

2.2 Statistical Consequences of a Change of Variables

In the development of regression equations to express in-use fuel usage in terms of EPA-rated fuel usage, it has been customary to employ least-squares techniques. The EPA-assigned number for a given vehicle class is regarded as a fixed quantity, and in-use deviations from this number are regarded as outcomes of a random variable having a distribution often referred to as the "error distribution." Conceivably the variance of in-use mpg or gpm values can be the same for all classes of vehicles. It seems more likely, however, that the variability of in-use fuel economy for a luxury automobile might be expected to be different from the variability of in-use fuel economy for a sub-compact vehicle. The assumption that variability is independent of the EPA-rating is essentially the statistical assumption of homoscedasticity. The assumption that variability is vehicle-class dependent is the assumption of heteroscedasticity. Under the latter assumption it is appropriate to employ a weighted least-squares procedure or to perform a transformation of variables aimed at "variance stabilization." Such a transformation tends to change a heteroscedastic error distribution into one which is essentially homoscedastic.

In the section on mathematical implications of a change of variables it was noted that a linear relation between in-use and EPA-rated fuel usage in mpg space tends to become nonlinear in gpm space. Similarly, if homoscedasticity applies in mpg space, it is most likely that the error distribution will be heteroscedastic in gpm space. Thus if the unweighted least squares method is employed to obtain linear regression equations, the error minimization process will be different in the two spaces. Other transformations can lead to still different distortions of the error distribution.

A good approximation to the effect of a transformation on the variance of a random variable is provided by the equation

$$\text{Var} [g(y)] \approx \text{Var} [Y] \cdot \{g'(E[Y])\}^2 \quad (18)$$

where g is the transforming function and g' is its derivative. For example, if $g(y) = 1/y \equiv z$, then $g'(y) = -1/y^2$ and

$$\text{Var } [Z] \approx \frac{\text{Var } [Y]}{\{ E [Y] \}^4} \quad (19)$$

or

$$\sigma_Z \approx \sigma_Y / \{ E [Y] \}^2.$$

Thus it is clear that if $\text{Var } [Y]$ is independent of its expected value, then $\text{Var } [1/Y]$ decreases monotonically as the expected value of Y increases.

For example, take $\text{Var } [Y] = 2$ mpg and examine the variance of $1/Y$ for various expected values of Y , as shown in Table 1.

Table 1 EFFECT OF A CHANGE OF VARIABLES ON ERROR VARIANCE

$E[Y]$ mpg	σ_Y (mpg)	$E[Z] = 1/E[Y]$ (gpm)	σ_Z (gpm)	$\sigma_Y/E[Y]$	σ_Z/Z
10	2	0.1	0.02	0.2	0.2
20	2	0.05	0.005	0.1	0.1
30	2	0.033	0.00222	0.0667	0.0667
40	2	0.025	0.00125	0.05	0.05

Mathematically,

$$\frac{\sigma_Z}{E[Z]} = \frac{\frac{\sigma_Y}{\{ E[Y] \}^2}}{\frac{1}{\{ E[Y] \}}} = \frac{\sigma_Y}{E[Y]}$$

and it is seen that the coefficient of variation is a function of mpg but is the same function in both spaces. Thus if one performs unweighted least squares in mpg space, all mpg values exert the same influence so far as error minimization is concerned. In gpm space, however, the regression equation will be influenced more strongly by low mpg vehicles than by high mpg vehicles.

In addition to the effect of a reciprocal transformation on variance stabilization there is an effect on the form of the error distribution itself (see Figure 1). A symmetric, normal distribution becomes appreciably skewed when subjected to a reciprocal transformation, and

FIGURE 1 Effect of a Reciprocal Transformation on a Normal Error Distribution

this skewness can affect any inferences based on a normal distribution. This observation highlights the need for concern for the form of the error distribution and whether there is a choice of metameter which tends to (1) yield a normal error distribution and (2) stabilizes variance over the range of expected values of fuel usage. It would be further desirable if this choice would also lead to unambiguous results independent of whether fuel usage is viewed in mpg space or gpm space. The logarithmic transformation appears to be such a happy choice, as will be discussed in Section 2.2.1. That discussion will be followed by a comparison of mpg space, gpm space, and log space in Section 2.2.2.

2.2.1 The Logarithmic Transformation

The thrust of this discussion is that it may be appropriate to compare in-use and EPA-based fuel-economy (fuel-consumption) by first performing a logarithmic transformation of these quantities. It will be shown that the logarithmic transformation has the desirable property of isomorphism, that it may be variance stabilizing, and that it may provide an error distribution which is compatible with sources of error attendant to fuel usage by automotive vehicles.

2.2.1.1 Mathematical Isomorphism

The mathematical property of isomorphism arises in the theory of groups, but its formal definition can be paraphrased to advantage in the context of fuel economy/fuel consumption. Basically, the notion is that a one-to-one correspondence is established between the elements in two spaces and that this correspondence is preserved under a certain mathematical operation on the elements in the two spaces.

Consider two real numbers N_1 and N_2 and their corresponding logarithms n_1 and n_2 . The product of N_1 and N_2 is a third real number N_3 , and the sum of the two numbers n_1 and n_2 is a third number n_3 . The property of isomorphism assures that N_3 and n_3 are corresponding elements, even though they arise from mathematical operations performed on N_1 and N_2 in the preimage space and n_1 and n_2 in the image space. This property has an interesting consequence when applied to a logarithmic transformation of either miles per gallon or gallons per mile.

Let x denote EPA miles per gallon and y denote in-use miles per gallon. Plot y against x in logarithmic coordinates and assume a linear relationship. Then

$$\log y = a' + b \log x \quad (20)$$

or, by antilog transformation,

$$y = a x^b \quad (21)$$

where $a = \log a'$.

Now take the inverse of (21) to obtain

$$\frac{1}{y} = \frac{1}{a x^b} = \frac{1}{a} x^{-b} \quad (22)$$

Then logarithmic transformation of (22) yields

$$\log (1/y) = \log (1/a) - b \log x \quad (23)$$

or

$$\begin{aligned} - \log y &= - \log a - b \log x \\ &= - a' - b \log x \end{aligned} \quad (24)$$

which, by change of sign, is identical to (20). Thus if a logarithmic transformation is performed, it makes no difference whether this transformation is applied to miles per gallon or gallons per mile, since the inverse transformation can be made in either space without the need to recompute the regression relation between in-use and EPA numbers.

A question, of course, arises as to whether an equation of the form of (21) is a "reasonable" model for either miles per gallon or gallons per mile, in view of the previous use of linear regression equations. In this connection, it is important to make two observations. First a linear regression equation is unacceptable as the EPA-rated fuel economy approaches zero, unless that equation has zero intercept. Such is not usually the case, so it must be assumed that the "correct"

relation is curvilinear. Second, within the usual range of miles per gallon, logarithmic transformation of a linear regression equation in mpg space introduces little nonlinear distortion. This fact can be illustrated by a simple example.

Take a "typical" linear relation in mpg space, such as

$$y = 2.5 + 0.7x \quad (25)$$

The computed y values for 5-mpg increments of x are shown in Table 2, with the corresponding logarithms of x and y .

Table 2 LOGARITHMIC TRANSFORMATION OF A LINEAR REGRESSION EQUATION

<u>x (mpg)</u>	<u>y (mpg)</u>	<u>log x</u>	<u>log y</u>
10	9.5	2.303	2.251
15	13.0	2.708	2.565
20	16.5	2.996	2.803
25	20.0	3.219	2.996
30	23.5	3.401	3.157
35	27.0	3.555	3.296

The extent of nonlinear curvature introduced by the transformation is shown in Figure 2.

FIGURE 2 Logarithmic Plot of In-Use vs. EPA Miles Per Gallon as Derived from a Linear Regression Equation

A straight line can be "forced" to fit the computed nonlinear relation in various ways--e.g., through two points or by least squares error minimization. Even if the line is put through the two extreme points, as shown in Figure 2, the maximum "lack-of-fit" error is no more than about 0.5 mpg. In short, the logarithmic transformation is not incompatible with the assumption of a linear relation as previously assumed in either mpg or gpm space, and may be more realistic as fuel-economy values approach zero.

2.2.1.2 Variance Stabilization

There is evidence that the variance of in-use fuel-economy values is dependent upon the expected or average value of fuel economy for a given class of vehicles. For example, vehicles averaging--say--30 miles per gallon tend to show more variability from case to case than do vehicles averaging--say--10 miles per gallon. This postulated greater variability could arise if the more fuel-efficient automobiles were more sensitive to changes in use scenarios, ambient conditions, state of vehicle maintenance, and the like. Evidence to this effect is provided by the fact that the absolute difference between urban and highway fuel economy values, as stated in the Fuel Economy Guide, tend to be greater for subcompact vehicles than for--say--luxury vehicles. More evidence pertinent to this assumption can be adduced by statistical analysis performed on homogeneous groups of vehicles.

As a reasonable hypothesis, let it be assumed that the variance among mpg values increases with average mpg. Specifically, let us assume a constant coefficient of variation--that is, the standard deviation of mpg values is a constant fraction of the mean mpg value. It will be seen that logarithmic transformation of mpg tends to be variance stabilizing.

To argue this point, let us return to equation (18):

$$\text{Var } [g(y)] \approx \text{Var } [Y] \cdot \{ g'(E[Y]) \}^2 \quad (18)$$

But

$$\text{Var } [Y] = k \{ E[Y] \}^2$$

and it is assumed that

$$g(y) = \log y.$$

Accordingly,

$$g'(E[Y]) = 1/E[Y]$$

and, upon substitution in (18),

$$\text{Var } [g(Y)] \approx k \{ E[Y] \}^2 \cdot (1/E[Y])^2 = k$$

Thus if the assumption of constant coefficient of variation is correct, an unweighted least squares solution can be used to develop the appropriate regression equation in log space.

2.2.1.3 Lognormal Error Distribution

A lognormal distribution is one in which the logarithm of the random variable has a normal distribution. It has been proposed as a model for the distribution of exhaust emissions on the basis of an argument based on multiplicative rather than additive combination of errors arising from various sources.* It is of interest to examine

* Automobile Exhaust Emission Surveillance--A Summary, EPA Report No. APTD-1544 (March 1973).

the possibility that such a distribution might apply to the errors involved in fitting a regression equation to fuel-economy data in log space.

Consider the case of n sources of multiplicative error--that is, consider n random variables $Z_1, Z_2, \dots, Z_n$ which combine according to the relation $Z = Z_1 Z_2 \dots Z_n$. Perform a logarithmic transformation to obtain

$$\log Z = \log Z_1 + \log Z_2 + \dots + \log Z_n$$

or

$$\theta = \theta_1 + \theta_2 + \dots + \theta_n$$

where $\theta = \log Z$ and $\theta_i = \log Z_i$. By the central limit theorem, θ tends to approach a normal distribution regardless of the distributions of $Z_1, Z_2, \dots, Z_n$, provided these distributions obey certain rather loose assumptions, such as having finite variance. By the isomorphism property of logarithms, however, the addition of error sources in log space translates into a multiplication of errors in mpg space.

Whether fuel-economy errors from various sources tend to combine multiplicatively is a question which requires further engineering analysis. It suffices, however, to note that as long as the coefficient of variation is relatively small there should be relatively little difference between inferences based on a lognormal assumption and those based on a normal assumption. This point can be made clear by an example.

Suppose that one is concerned with a collection of vehicles for which the mean value of fuel economy is 20 mpg and the standard deviation is 2 mpg. Under the assumption of normality, 68.3% of all cases would lie within the range from 18 to 22 mpg, 95.5% within the range from 16 to 24 mpg and 99.7% within the range from 14 to 26 mpg. Under the lognormal assumption, the corresponding limits are expressed in terms of the median value of fuel economy and an appropriate multiplicative factor. For a median of 20 mpg and a coefficient of variation of 0.1 (corresponding to $\sigma = 2$ mpg), 68.3% of all cases would lie within the interval between

$$\frac{20}{1.1} = 18.2 \text{ mpg and } (20)(1.1) = 22 \text{ mpg.}$$

Further, 95.5% of all cases would lie within the interval between

$$\frac{20}{(1.1)^2} = 16.5 \text{ mpg and } (20)(1.1)^2 = 24.2 \text{ mpg,}$$

and 99.7% within the interval between

$$\frac{20}{(1.1)^3} = 15.0 \text{ mpg and } (20)(1.1)^3 = 26.6 \text{ mpg.}$$

Note that these bounds do not differ to any large extent from those derived from the normality assumption.

As a final observation it is pointed out that when vehicles are aggregated according to some rational design classification such as weight or interior volume, the mean fuel economy for vehicle models tends to vary about the overall mean for the given weight or volume class. Now recall that other sources of variation are such that the standard deviation within each model is proportional to the mean for that model. The result is that the within-model and between-model sources of variability tend to combine multiplicatively as previously postulated.

2.2.1.4 Recapitulation

In conclusion, the assumption of a geometric or power-law relation between in-use and EPA-based fuel usage is attractive on several bases. First, it can be expressed as a linear relation in log space and the resulting regression equation gives results which can be transformed to consistent values of either miles per gallon or gallons per mile. Second, the logarithmic transformation is variance stabilizing under the assumption that within-class variance is proportional to the class mean. Finally, the transformation is not expected to produce a serious distortion of the error distribution and may, in fact, have the effect of making it conform more closely to the normal distribution of errors.

2.2.2 Change of Variables: Some Numerical Comparisons

To illustrate and make concrete some of the analytical contentions of this paper, the following discussion presents some numerical examples based on both synthesized and actual fuel-usage data. The synthetic data has the advantage that the assumptions made in the analytical development can be rigorously implemented. The actual fuel-usage data, though less subject to control of assumptions, has the added element of realism.

2.2.2.1 Comparison Based on Synthetic Data

In earlier sections of this paper three possible choices of metameter for expressing vehicle fuel-usage characteristics have been advanced: miles per gallon, gallons per mile, and the logarithm of either mpg or gpm. Further, various assumptions have been made with regard to error distributions affecting the development of regression equations, notably whether the variance of the error distribution can be considered to be independent of vehicle-class mean (homoscedastic) or class mean dependent (heteroscedastic). In the following discussion material examples are presented to illustrate the effect of these various assumptions on the resulting regression equation and its predictive capability.

First, let us consider the data-generation process involved in the simulation exercise. A specific relation between EPA-based miles per gallon (x) and in-use miles per gallon (y) was assumed:

$$y = 2.5 + 0.7x \quad (26)$$

It was further assumed that x was not subject to error, but that y had a specified error distribution. In all instances this distribution was assumed to be normal, but in one case the standard deviation was assumed to be constant for all values of x and in the other case was assumed to be proportional to x^* . For the homoscedastic case it was assumed that $\sigma_y = 2$ mpg. For the heteroscedastic case it was assumed that $\sigma_{y|x} = k E[Y|x]$, with $k = 0.1$.

* Note that, by virtue of (26), these assumptions are equivalent to the assumptions of being independent of or dependent on y .

To prepare the various data bases, equation (26) was used to generate values of y for $x = 10, 15, 20, 25, 30, 35$ and 40 mph. Then, each of these computed values was "contaminated" with errors produced by an appropriate random number generation process. Seven y -values were generated for each x -value, and these 49 data points were employed in the regression analyses. The gpm data were obtained simply by taking the reciprocals of the generated y -values or their reciprocals.

Table 3 gives the regression equations obtained for the homoscedastic and heteroscedastic cases by the unweighted least-squares approach. Figures 3 and 4 display the corresponding curves.

Table 3 SUMMARY OF REGRESSION EQUATIONS
FOR HYPOTHESIZED $y = 2.5 + 0.7x$

Case Ident.	Dependent Metameter	Error Parameter	Intercept a	Slope b	r
1	y	$\sigma_y = 2.0$	1.514	0.734	0.959
2	$1/y$	$\sigma_y = 2.0$	-0.000827	1.282	0.900
3	$\log y$	$\sigma_y = 2.0$	-0.0708	0.950	0.946
4	$\log (1/y)$	$\sigma_y = 2.0$	0.0708	0.950	0.946
5	y	$\sigma_y = 0.1 E[Y]$	1.926	0.719	0.954
6	$1/y$	$\sigma_y = 0.1 E[Y]$	0.00816	1.049	0.964
7	$\log y$	$\sigma_y = 0.1 E[Y]$	0.1596	0.880	0.964
8	$\log (1/y)$	$\sigma_y = 0.1 E[Y]$	-0.1596	0.880	0.964

Note that all of the regressions show a relatively high coefficient of correlation as based on the actual metameter used in curve fitting. Note, also, that although the various curves are not grossly different for the several cases, they do reflect the weighting process attendant to the transformations involved.

σ -constant

FIGURE 3 Regression Relations for Homoscedastic Error Distributions

σ -proportional

FIGURE 4 Regression Equations for Heteroscedastic Error Distributions

Table 4 provides a comparison of the predictive capabilities for the various curve-fitting processes and makes quantitative the differences observed graphically in the figures. The values tabulated are obtained from the relationships of Table 3 as identified by case number. It is of interest to note that all the curves are in close agreement near the center of the range of x covered and tend to diverge at the low and high extremes of x . Extrapolation beyond these points would, accordingly, lead to even further divergence, and it is for this reason that the choice of regression model could affect seriously the nature of any inferences made on the basis of such extrapolation.

Table 4 COMPARISON OF PREDICTIVE CAPABILITIES OF REGRESSION EQUATIONS

x (mpg)	Y (mpg)							
	Case: 1	2	3	4	5	6	7	8
	σ_y Constant				σ_y Variable			
10	8.9	7.8	8.3	8.3	9.1	8.8	8.9	8.9
15	12.5	11.8	12.2	12.2	12.7	12.8	12.7	12.7
20	16.2	15.8	16.0	16.0	16.3	16.5	16.4	16.4
25	20.0	19.8	19.8	19.8	19.9	20.0	20.0	20.0
30	23.5	23.9	23.6	23.6	23.5	23.2	23.4	23.4
35	27.2	27.9	27.3	27.3	27.1	26.2	26.8	26.8
40	30.9	32.0	31.0	31.0	30.7	29.1	30.2	30.2

2.2.2.2 Comparison Based on Actual Fuel-Economy Data

In the previous examples the nature of the error distribution was known. Such is not the case when dealing with actual, fuel-economy data. To obtain an appreciation of how regression lines differ for real data when developed in mpg and gpm space, a set of 50 pairs of measurements were analyzed. These data are tabulated in Appendix II.

The equations developed are as follows:

$$\begin{aligned} \text{Fuel-economy (mpg) space:} \\ y = 3.35 + 0.647 x \end{aligned} \quad (27)$$

$$\begin{aligned} \text{Fuel-consumption (gpm) space:} \\ 1/y = 0.009 + 1.045(1/x) \end{aligned} \quad (28)$$

Plots of these equations are given in Figure 5.

The curves cross at the points

$$\begin{array}{ll} \text{and} & \begin{array}{ll} x = 31.13, & y = 23.49 \\ & x = 19.31, & y = 15.84 \end{array} \end{array}$$

and accordingly yield identical results at two values of x , namely 31.13 and 19.13.

FIGURE 5 Fuel-Economy (mpg) and Fuel-Consumption (gpm) Bases of Comparing EPA and In-Use Data

Comparisons at other values of x are given in Table 5.

Table 5 COMPARISON OF REGRESSION EQUATION PREDICTIONS
IN MPG AND GPM SPACE.

x (mpg)	y (mpg)	
	By Eq. (27)	By Eq. (28)
5	6.6	4.6
10	9.8	8.8
15	13.1	12.7
(19.31)	(15.84)	(15.84)
20	16.3	16.3
25	19.5	19.7
30	22.8	22.8
(31.13)	(23.49)	(23.49)
40	29.2	28.5

The most notable difference between the two relations is that they tend to diverge as they approach the origin. This divergence results partly from the different bases of error minimization, but mostly from the fact that the gpm relation constrains the curve to go through the origin when inverted and plotted in mpg space. The direct solution, on the other hand, yields a finite intercept not capable of being reconciled with actual vehicle performance. By way of comparison it should be noted that the gpm intercept in eq. (28) implies a limiting realizable fuel economy of $1/(0.009) = 111$ mpg. Whether or not this limiting value is realistic, the notion of an upper limit for fuel economy is intellectually acceptable and might be regarded as a goal for projections of the technology represented by the data base.

3. Conclusions

The prospect of obtaining different representations of the relation between in-use and EPA-based fuel economy places a severe onus on any analyst attempting to evaluate the so-called fuel-economy shortfall. Two sources of difficulty have been identified: mathematical and statistical.

In regard to the mathematical source of difficulty, it must be borne in mind that the principal purpose for developing an empirical relation between EPA-based and use-based numbers is to be able to "predict" in-use values for cases for which no data are available. It has been shown in this paper that several regression models--miles-per-gallon, gallons-per-mile, log (miles-per gallon)--tend to give substantially equivalent results within the interpolation range of the available data. When the relations are extrapolated beyond the range of the data, however, significant divergence occurs among the models as a result of differences in the curvatures of the regression functions. In the paper reasons are given to support the belief that the "true" relation between EPA-based and in-use values is nonlinear rather than linear when considered over the entire range of possible fuel-economy (fuel-consumption) values.

In regard to the statistical source of difficulty, it is pointed out that each choice of model has an associated and implicitly assumed error distribution. This error distribution interacts with the least-squares error-minimization procedure in such a way as to alter error weightings as one transforms from--say--miles-per-gallon to gallons-per-mile or log (miles-per-gallon). The resulting change in error weighting has an effect on the estimation of the model parameters, but there is reason to believe that this effect tends to be small.

In conclusion, the empirical basis for representing the relation between in-use and EPA-based fuel usage should be chosen in such a way that it is compatible with engineering facts and with the facts of error distributions affecting in-use fuel economy (consumption). It is believed that either gallons-per-mile or log (miles-per-gallon) is preferable to miles-per-gallon as a basis for developing a predictive relation between in-use and EPA-based numbers.

APPENDIX I

DERIVATION OF FUEL-ECONOMY (mpg) REGRESSIONS CORRESPONDING TO CONSTANT DIFFERENCE BETWEEN IN-USE AND EPA FUEL CONSUMPTION (gpm)

If it is assumed that the relation between in-use and EPA fuel economy is linear in gallons-per-mile space, then

$$y = a + bx \quad (I-1)$$

where

x = EPA-rated fuel economy (gpm)

and

y = In-use fuel economy (gpm)

Let $X = 1/x$ and $Y = 1/y$, where X and Y denote, respectively, EPA-based and in-use fuel consumption in gallons-per-mile. Then $x = 1/X$ and $y = 1/Y$ and, upon substitution in (I-1),

$$Y = \frac{1}{a + b(1/X)}$$

or

(I-2)

$$Y = \frac{X}{b + aX}$$

Now suppose that it is postulated that for some range of values of X (and hence of x), the difference between Y and X is some constant,

call it k . Then:

$$Y - X = k$$

or

$$\frac{X}{b + aX} - X = k \quad (I-3)$$

From (I-3) one can constrain X to two values to obtain two equations which can be solved for a and b in terms of k . Thus, the constant difference would hold at the end-points of a range of X values, and presumably would hold approximately over that interval. Operating on (I-3), one obtains

$$\left. \begin{aligned} X - (b + aX) X &= k(b + aX) \\ X - bX - aX^2 &= kb + ak X \\ bX + bk + ak X + aX^2 &= X \\ b(X + k) + a(kX + X^2) &= X \end{aligned} \right\} \quad (I-4)$$

Let

$$B = X + k$$

$$A = kX + X^2$$

Then we can write (I-4) simply as

$$Bb + Aa = X$$

and, by using two values of A and B obtain the equations

$$\left. \begin{aligned} B_1 b + A_1 a &= X_1 \\ B_2 b + A_2 a &= X_2 \end{aligned} \right\} \quad (I-5)$$

where

$$B_1 = X_1 + k$$

$$B_2 = X_2 + k$$

$$A_1 = k X_1 + X_1^2$$

$$A_2 = k X_2 + X_2^2$$

Writing (I-5) in matrix notation:

$$\begin{pmatrix} B_1 & A_1 \\ B_2 & A_2 \end{pmatrix} \begin{pmatrix} b \\ a \end{pmatrix} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$$

$$\begin{pmatrix} b \\ a \end{pmatrix} = \frac{1}{A_2 B_1 - A_1 B_2} \begin{pmatrix} A_2 & -A_1 \\ -B_2 & B_1 \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$$

or

$$b = \frac{A_2 X_1 - A_1 X_2}{A_2 B_1 - A_1 B_2}$$

$$a = \frac{B_1 X_1 - B_2 X_2}{A_2 B_1 - A_1 B_2}$$

APPENDIX II

ACTUAL DATA FOR COMPARISON OF FUEL-ECONOMY (mpg) AND FUEL-CONSUMPTION (gpm) BASES OF SHORT-FALL EVALUATION

<u>x</u> (EPA-based mpg)	<u>y</u> (In-use mpg)	<u>x</u> (EPA-based mpg)	<u>y</u> (In-use mpg)
27.0	19.9	19.9	16.1
20.0	13.3	16.9	15.6
17.9	15.9	17.9	13.7
20.3	18.1	26.3	20.0
16.2	13.6	27.0	20.7
14.9	13.7	19.3	16.3
17.9	15.0	20.6	17.9
15.2	13.2	18.2	15.3
19.3	16.3	16.6	15.6
27.4	21.6	17.9	15.7
20.6	17.0	18.5	14.4
17.9	14.8	14.9	12.5
18.2	15.5	17.2	11.8
17.9	15.2	17.2	17.2
18.2	14.9	15.6	13.6
19.3	15.1	18.9	16.4
26.6	18.3	17.9	14.9
27.0	19.0	15.2	13.0
15.9	13.6	17.9	14.3
14.9	11.5	17.2	14.8
14.9	11.7		
20.0	16.3		
18.9	17.6		
17.2	16.1		
26.3	20.2		
30.0	21.8		
30.0	25.5		
21.6	19.7		
17.9	14.1		
14.9	10.5

Interpretation Paper No. 2
Project No. 8431
US Dept. of Energy
August 1, 1979
H. T. McAdams

WEIGHTING-FACTOR CONSIDERATIONS IN THE REGRESSION
ANALYSIS OF VEHICLE FUEL USAGE

1. Introduction

Measures of in-use fuel consumption of automotive vehicles are acquired under conditions quite different from those affecting EPA-based measures. Whereas each vehicle model/engine number combination has an EPA miles-per-gallon figure quoted in the Mileage Guide, in-use data is usually available only for a subset of these vehicles. The overall picture of in-use vehicle fuel requirements is incomplete, therefore, and a statistical analyst must resort to techniques by which "missing points" in the data base can be filled in most logically and with the greatest possible credibility. To do so requires, in effect, a willingness to inject additional information in some way. This information augmentation usually takes the form of a "model" by means of which one can interpolate or extrapolate in-use fuel economy values, given the corresponding EPA-based values. As pointed out in Interpretation Paper No. 1,* the choice of model can have a significant effect on the conclusions drawn from the limited data available.

*H. T. McAdams, "Implications of Change of Variables in the Regression Analysis of Vehicle Fuel Usage," Interpretation Paper No. 1 (June 8, 1979).

It should also be clear that even with an optimum model, the extent and quality of data is a consideration which affects conclusions. If one has control of the data-acquisition process, there exist well-formulated principles of experiment design by which data collection can be optimized. This optimization consists, for the most part, of assigning appropriate weights to the various items of data with due regard to the purpose and intent of the statistical estimation process. When one has less than full control of data acquisition, a certain amount of compromise, with optimality is inescapable. Within the constraints of available data, however, it is possible to seek the best suboptimal approach to its analysis, to particularize the consequences of any compromises involved, and to quantify the bounds within which conclusions are applicable. It is to these considerations that this Application Paper is addressed.

2. Forms of Data Weighting

There are at least three different ways in which data weighting can affect the estimation of fuel-economy measures. First, there is a weighting which characterizes the relative abundance of vehicles of various makes, models and engine numbers in the vehicle population. The weights involved are those on which a sales-weighted average fuel-requirement number is based. Second, there is a weighting which characterizes the relative abundance of vehicles of various makes, models and engine numbers in the sample of vehicles available in the data base. The weights involved are not necessarily in consonance with the population weights, though they could be made so if one were in a position to propose, in advance of data collection, how the sample were to be taken. Finally, there are weights which can be assigned, for purposes of calculation, to the various makes, models and engine numbers at the option of the analyst after the fact of data acquisition. Again, these weights do not need to be the same as either the population weights or the sample weights if they can accomplish some desirable statistical end in the analysis.

Quite clearly, these three types of weighting are interrelated. Their effect on the estimation of in-use fuel economy and the shortfall of in-use numbers relative to EPA-based numbers will be discussed below.

2.1 Population weights

Consider N classes of vehicles, each class being defined to the level of a line entry in the Mileage Guide. Let these classes be denoted $V_1, V_2, \dots, V_N$. For each class V_i there is an associated EPA based fuel-requirement number r_i , where we eschew for the moment whether this number is expressed as miles per gallon (mpg) or gallons per mile (gpm). Also for each class there is a sales-weight fraction f_i which expresses the ratio of the number of vehicles in class V_i to the number of vehicles in all the classes. Then

$$\rho = \sum_{i=1}^N f_i r_i \quad (1)$$

in the sales-weighted average of all EPA fuel-requirement ratings.*

* Note that if the r_i are in terms of gpm, ρ is a sales-weighted gpm which can be inverted to obtain the sales-weighted harmonic mean mpg.

So far as vehicle certification is concerned, equation (1) contains no random variables. The r_i are presumed to be fixed and the f_i are presumed to be known from actual sales records or other relevant sources of information. Let us now consider, however, the in-use fuel-requirement picture. Each vehicle in class V_i is unique, both with regard to production variation from the test vehicle and with regard to the manner in which the vehicle is driven and maintained. Thus if the population contains n_i vehicles in class V_i , there will be as many values of fuel economy or fuel consumption,* and these values will have a statistical distribution with a corresponding mean μ_i and variance σ_i^2 .

Our concern is to determine a relation between the certification and in-use fuel requirements of vehicles and to what extent there is a shortfall. Inasmuch as the extent of the shortfall may be vehicle-class dependent, interest centers on a scheme for evaluating the shortfall on a class-by-class basis. A complete solution would require determining the distribution parameters for each V_i , though it might suffice for most purposes if one knew the μ_i and σ_i^2 . A sales-weighted average in-use number could then be computed as

$$\mu = \sum_{i=1}^N f_i \mu_i \quad (2)$$

and a sales-weighted shortfall d as

$$d = \sum_{i=1}^N f_i (\mu_i - \rho_i) \quad (3)$$

In reality, the μ_i are unknown because it is impossible to know the fuel performance of every vehicle in the population. Rather, the μ_i must be estimated from a sample of the population, and a major concern is with an optimum scheme for making such estimates within specified resource constraints.

* In reality a single value for each vehicle is itself a simplification, since fuel consumption varies continuously with time and can be reduced to a single value only by some averaging process.

2.2 Sample weights

Each of the vehicle classes $V_1, V_2, \dots, V_N$ consists of a finite number of vehicles and these vehicles comprise the population of interest. The fact that the population is finite need not concern us at the moment, though in reality there are distinctions between finite and infinite sample spaces. What is of very important concern, however, is that the sample will contain only n_1 vehicles from class V_1 , n_2 vehicles from class V_2 , and so on and that the n_i can vary from zero to the total number of vehicles in the class. Usually the n_i will be very small relative to the total number of vehicles in V_i and, by virtue of sampling variation, the estimates $\bar{x}_i$ of the μ_i will be subject to sampling error. Moreover, it may well be that many of the vehicle classes V_i may be totally unrepresented in the sample. In such instances one must resort to "cross-class" comparisons in order to infer the class means, and it is for this purpose that the regression approach is useful.

Suppose that one is limited to K vehicles in the overall sample but is at liberty to allocate these samples arbitrarily among the N vehicle classes $V_1, V_2, \dots, V_N$. Suppose also that the sales-weight fractions $f_1, f_2, \dots, f_N$ for the population are known. How should the K vehicles be assigned in order to achieve "best estimation" of in-use fuel requirements in some sense?

Two important questions arise at this point. First is the matter of whether it is desired to estimate the separate μ_i or whether one is content to estimate some function of the μ_i , such as the sales-weighted average

$$\mu = f_1 \mu_1 + f_2 \mu_2 + \dots + f_N \mu_N.$$

A corollary to this question is whether it is desired that the μ_i be estimated with equal precision or with a precision which is more critical for some classes than others. Second is the question of whether the σ_i^2 are the same for all classes or vary in magnitude from class to class. How these questions are answered has a very important bearing on what constitutes an optimum sampling procedure and associated estimation technique.

To illustrate this point it will simplify matters to assume that $\sigma_1^2 = \sigma_2^2 = \dots = \sigma_N^2$. In this case it is clear that if one wishes to estimate the μ_i with equal precision, then it is imperative that the K observations be divided equally among the N vehicle classes $V_1, V_2, \dots, V_N$. Suppose, however, that one wishes to obtain a minimum variance estimate of the sales-weighted mean

$$\mu = f_1 \mu_1 + f_2 \mu_2 + \dots + f_N \mu_N.$$

Now it should be intuitively evident that if one of the f_i dominates, the corresponding vehicle class V_i exerts the greatest influence on μ and should accordingly be sampled more extensively than the other vehicle classes. Indeed, it turns out that the optimum allocation of observations is that the i th vehicle class V_i is assigned $f_i K$ observations. In other words, the relative proportion of vehicles in the sample should be the same as in the population. Formal justification for this assertion is provided in Appendix I.

It goes without saying that one can not have the "best of both worlds"--that is, one can not have a sample which is best for estimating the sales-weighted μ and the individual μ_i unless $f_1 = f_2 = \dots = f_N$. Moreover, if the sample weighting are fixed as a result of data-acquisition constraints, they most likely do not give rise to optimum estimation of either μ or the μ_i . The extent to which the sample weights are less than optimum depends on the extent to which they differ from the requirements specified above.

3. Analyst-Imposed Weighting Factors

Suppose that the available data base does not satisfy requirements for optimum estimation of either the sales-weighted mean or the individual class means. There is no reason ab initio that the statistical analyst can not impose weightings of his own choice in an attempt to compensate for lack of optimality or for other reasons. As will be seen, however, such weightings can not be assigned arbitrarily. In his choice of weighting factors, the analyst may be forced to make a choice between conflicting estimation requirements, such as unbiasedness and minimum variance, inasmuch as the limitations of the data may be such as to preclude the realization of both.

This point is best illustrated by an example. Suppose that the sample contains equal numbers of observations for each of the vehicles classes $V_1, V_2, \dots, V_N$ but one wishes to estimate the sales-weighted average μ from the sample means $\bar{x}_i$. Is it better to assign each $\bar{x}_i$ the same weight or the corresponding sales weight f_i ? Or is there some other set of weights $g_1, g_2, \dots, g_N$ preferable to either of the above options? As will be evident in what follows, the answer to these questions depends on the properties desired in the estimate.

If there are K observations equally divided among N vehicles classes, then each class contains K/N observations and each of the class sample means $\bar{x}_i$ has variance

$$\frac{\sigma^2}{K/N} = \frac{N\sigma^2}{K}$$

Assume that weightings g_i are to be applied to the $\bar{x}_i$ to obtain an estimate of the weighted population mean

$$\mu = f_1 \mu_1 + f_2 \mu_2 + \dots + f_N \mu_N. \quad (4)$$

It is of interest to select the g_i so as to obtain a minimum-variance estimator and preserve, if possible, the property of unbiasedness. If it is assumed that the sample means are independent, one can write

$$\begin{aligned} \text{Var} [\hat{\mu}] &= g_1^2 \left(\frac{N\sigma^2}{K} \right) + g_2^2 \left(\frac{N\sigma^2}{K} \right) + \dots + g_N^2 \left(\frac{N\sigma^2}{K} \right) \\ &= \frac{N\sigma^2}{K} \sum_{i=1}^N g_i^2 \end{aligned} \quad (5)$$

It is desired, then, to assign the g_i , subject to the restriction

$$\sum_{i=1}^N g_i = 1,$$

so that (5) is minimized.

Write

$$g_N = 1 - \sum_{i=1}^{N-1} g_i$$

and rewrite (5) as

$$\text{Var} [\hat{\mu}] = \frac{N\sigma^2}{K} \left[\sum_{i=1}^{N-1} g_i^2 + \left(1 - \sum_{i=1}^{N-1} g_i \right)^2 \right] \equiv G. \quad (6)$$

Considering G as a function of the g_i , one can minimize (6) as follows:

$$\frac{\partial G}{\partial g_i} = \frac{N\sigma^2}{K} \left[2 g_i - 2 \left(1 - \sum_{i=1}^{N-1} g_i \right) \right] = 0$$

$$g_i = 1 - \sum_{i=1}^{N-1} g_i = g_n$$

for all i . Therefore, (5) is minimized when $g_1 = g_2 = \dots = g_n$.

Note that the above conclusion is tantamount to the assertion that

$$\hat{\mu} = g_1 \bar{x}_1 + g_2 \bar{x}_2 + \dots + g_N \bar{x}_N \quad (7)$$

Taking expectations in (6) one has

$$\begin{aligned} E [\hat{\mu}] &= E [g_1 \bar{x}_1 + g_2 \bar{x}_2 + \dots + g_N \bar{x}_N] \\ &= g_1 E [\bar{x}_1] + g_2 E [\bar{x}_2] + \dots + g_N E [\bar{x}_N]. \end{aligned}$$

Since the class sample mean $\bar{x}_i$ is an unbiased estimator of the class population mean, $E [\bar{x}_i] = \mu_i$ by definition and

$$E [\hat{\mu}] = g_1 \mu_1 + g_2 \mu_2 + \dots + g_N \mu_N \quad (8)$$

Note that (8) provides an unbiased estimator of μ only if

$$\mu = g_1 \mu_1 + g_2 \mu_2 + \dots + g_N \mu_N$$

--that is, only if the weights assigned to the sample means are the same as those which are inherent in the definition of the weighted population mean as given in (4). Thus if equal weights are assigned to the sample means, the resulting estimator of μ is biased unless all vehicle classes are equally weighted in the population. This conclusion does not preclude the fact that the estimator based on equal weightings of the class means has smaller variance than the estimator based on unequal weightings. Note, however, that this statement applies only if each of the class sample means is an average of the same number of observations.

Under ideal circumstances one seeks the best linear unbiased estimator (BLUE) of μ , where "best" implies the estimator with minimum variance. As shown in Appendix I, minimum variance is achieved by designing the sample so that observations are allocated to the vehicle classes in accordance with the population weightings. Since these are exactly the weighting which must be applied to the sample means in order to achieve unbiasedness, the conditions of BLUE can be satisfied only if the sample is properly configured at the outset. If such is not the case, the statistical analyst must be satisfied with less than ideal estimation and must defend his analytical approach according to the practical requirements of the problem.

4. Data Lacunae and Regression Analysis

The weighting considerations discussed above apply to circumstances in which there exists an estimate of the vehicle class mean $\bar{x}_i$ for every vehicle class V_i , even though this estimate may be based on minimal data. In short, it is assumed that each and every V_i is represented in the data base, if only by one vehicle. In actuality, however, there may be many vehicle classes for which no in-use fuel-consumption data is available. Unless these data gaps can be filled by some means, there is no hope of obtaining an estimate of the sales-weighted, in-use, mean fuel economy or fuel-economy shortfall.

Regression analysis provides a means for filling these lacunae, and its application is subject to the same weighting considerations as discussed above. The implications of these considerations in the context of regression analysis is the subject treated in the following discussion.

4.1 The Regression Model: Some Basic Considerations

In a certain sense, regression analysis is a generalization of the process by which a class population mean is estimated by the corresponding class sample mean. Both are based on least-squares minimization of errors. To clarify this point it will be informative to review the statistical basis of the commonly used sample average as a prelude to a discussion of the regression model.

Consider a class of vehicles V and suppose that this class has a population mean fuel economy of μ miles per gallon. Suppose, further, that one has a random sample of n vehicles from the class and for each an in-use fuel-economy value. What is the basis for employing

$$\bar{x} \doteq \frac{1}{n} \sum_{i=1}^N x_i,$$

where the x_i are the sample observations, as an estimate of μ ?

Each of the observations x_i can be expressed as

$$x_i = \mu + e_i$$

where e_i is a deviation from the true mean. It is assumed that e_i is the outcome of an error random variable ε_i having the properties

$$E(\epsilon_i) = 0$$

$$E(\epsilon_i^2) = \sigma^2$$

$$E(\epsilon_i \epsilon_j) = 0, i \neq j$$

Then, upon rearrangement, one has

$$e_i = x_i - \mu$$

Consider the sum of the squared errors as a function of μ , so that

$$F(\mu) = \sum_{i=1}^N e_i^2 = \sum_{i=1}^N (x_i - \mu)^2$$

Let us choose that value $\hat{\mu}$ of the variable μ which minimizes $F(\mu)$.

$$\frac{\partial F(\mu)}{\partial \mu} = -2 \sum_{i=1}^N (x_i - \hat{\mu}) = 0$$

$$N\hat{\mu} = \sum_{i=1}^N x_i$$

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^N x_i \equiv \bar{x}$$

Thus, the sample mean $\bar{x}$, as customarily defined, is the least-squares estimator of the population mean μ .

With regard to regression analysis, there is a direct parallel in the fact that each of the observations y_i of the dependent variable is expressed as

$$y_i = \beta_0 + \beta_1 x_i + e_i$$

If V_1 and V_3 are considered separately, the best estimator of their in-use fuel economies are their respective sample means $\bar{y}_1 = 9$ and $\bar{y}_2 = 23$. If a linear regression equation is fitted to the five data points, one obtains

$$y = 2.0 + 0.7x$$

For $x = 10$, $y = 9.0$ and for $x = 30$, $y = 23$. Thus, in the particular case cited, the same estimates are obtained by linear regression as by class-by-class averaging. If the proposed linear relation is really correct, then one can postulate

$$y = 2.0 + 0.7(22) = 17.4$$

as the estimated in-use fuel economy for class V_2 . Indeed, it can be argued that 17.4 is a better estimate for V_2 than 9.0 is for V_1 or 23.0 is for V_3 . The argument rests strongly on our belief in the validity of the linear relation which, in effect, draws information from all five observations in the estimation for V_2 , whereas $\bar{x}_1$ draws information from only two observations and $\bar{x}_3$ from only three.

Now it is true that Figure 1 presents an oversimplified picture in that there is no possibility of "model lack of fit," since only two values along the x-axis were involved. Similar arguments apply in the more general case, however, as discussed in Appendix II, where a formal development is presented to show how the allocation of data points affects the precision with which in-use fuel economy can be estimated for various EPA-based numbers.

To illustrate some of the consequences of the formal argument, a further simple example is in order. In this second example, let it be supposed that there are four vehicle classes, as follows.

V_1	7 Observations	$x = 15$
V_2	2 Observations	$x = 20$
V_3	0 Observations	$x = 30$
V_4	1 Observation	$x = 35$

A plot of the data is shown in Figure 2.

and one seeks to minimize the squared error function

$$\begin{aligned} F(\beta_0, \beta_1) &= \sum_{i=1}^N e_i^2 \\ &= \sum_{i=1}^N (y_i - \beta_0 - \beta_1 x_i)^2 \end{aligned}$$

The solution, previously presented in Application Paper No. 2,* essentially provides an estimate of the expected value (in a sense, the "population mean") of the dependent variable y as a continuous function of the independent variable x .

Suppose that one has three vehicle classes V_1 , V_2 , and V_3 , with $n_1 > 0$ observations of in-use fuel economy for V_1 and $n_3 > 0$ observations of in-use fuel economy for V_3 , but no observations for V_2 . Let it further be supposed, however, that the EPA-based numbers are known for V_1 , V_2 , V_3 . How can the available information be used to obtain an in-use fuel economy for V_2 ? An exemplification of the above hypothesis is shown in Figure 1, with $n_1 = 2$ and $n_2 = 3$.

*H. T. McAdams, Fuel Economy of Automotive Vehicles: A Compendium of Statistical Applications.

FIGURE 1: Hypothetical Regression Model

When evaluated for the x-values corresponding to V_1 , V_2 , V_3 and V_4 , the equation yields:

V_1	$x_1 = 15$	$y_c = 14.2$	$\bar{y}_1 = 14.0$
V_2	$x_2 = 20$	$y_c = 18.7$	$\bar{y}_2 = 19.5$
V_3	$x_3 = 30$	$y_c = 27.8$	Not known
V_4	$x_4 = 35$	$y_c = 32.4$	$\bar{y}_4 = 32.0$

In this case the class sample means do not agree with the values computed by the regression equation, but it is presumed, on the basis of the assumed linear relation, that the values calculated from the regression equation (the y_c values) may, in some cases at least, be better estimates of the population class means than are the sample averages. A more complete discussion of this matter is given in Appendix II.

The precision with which the regression equation estimates the expected value of y for a given value of x is a function of x as well as a function of the points in the original data. As shown in Appendix II, the variance of $\hat{y}$ for a given value of x is

$$\text{Var} [\hat{y}(x)] = \left[\frac{1}{N} + \frac{(x - \bar{x})^2}{\sum_{i=1}^N (x_i - \bar{x})^2} \right] \sigma^2$$

where σ^2 is the error variance and is usually estimated from the residuals. The expression within square brackets can be thought of as a multiplier which expresses the dependence of $\text{Var} [\hat{y}(x)]$ on x as well as on the data x_i , $i = 1, 2, \dots, N$. Note that the more widely spaced the x_i from the mean, the smaller is the variance of $\hat{y}$ for any position along the x-axis. Moreover, for a given set of data points, $\text{Var} [\hat{y}(x)]$ decreases as this position approaches the mean of the data x values. At $x = \bar{x}$, the variance is simply σ^2/N and is the same as if all the measurements had been made at $\bar{x}$.

With regard to weighting considerations, it is apparent that the regression equation, as computed, gives most weight to the data for V_1 and least weight to the data for V_4 . This conclusion follows from the

FIGURE 2: Hypothetical Regression Model

A solution of the least-squares regression equation gives

$$y = 16.9 + 0.911 (x - \bar{x})$$

or

$$y = 0.5 + 0.911 x$$

fact that each individual observation is given the same weight, and V_1 supplies most of the observations. By averaging the observations in each class and fitting the regression equation to these averages one can give each of the three classes of vehicles equal weight. This approach tends to bias the regression equation, however, and it is difficult to estimate the precision with which missing in-use values are estimated.

In the example cited, $\bar{x} = 18$. The average value of x lies close to $x = 15$, the point at which 7 out of the 10 observations were made. With regard to weighting considerations, it is apparent that the regression equation, as computed, gives most weight to the data for V_1 and least weight to the data for V_4 . One is tempted to attribute to that fact the fact that best estimation is achieved near the point at which most of the observations are concentrated. Such is not the case, however, because minimum variance would be achieved at $x = 18$ for any selection of points for which $\bar{x} = 18$. Indeed, as shown in Appendix II, there is an advantage, if one can do so, in selecting x -values so that they cluster at the two ends of the range of interest if one desires to estimate y as uniformly as possible at all points in the range. If one is interested only in the value of y corresponding to some (weighted) average of the x -values, however, it is important only that the actual average of the x_i is such that it conforms as nearly as possible to the weighted average of interest.

5. Implications of Estimation Theory for Fuel-Economy Assessment

Regression analysis with its attendant estimation theory has some interesting and curious implications in the context of fuel economy (fuel consumption) assessment and the estimation of shortfall. Of especial interest is the distinction between shortfall for specific vehicle species as opposed to a sales-weighted average shortfall. Also of interest is the extent to which the regression approach accommodates and reflects technological and design innovations.

First it is in order to exhibit an important consequence of linearity in the development of a regression equation, whether this equation is developed in mpg or gpm space. Suppose that a regression equation

$$y = b_0 + b_1x$$

has been developed and is to be used to estimate the average in-use fuel economy (fuel consumption) number corresponding to a sales-weighted average EPA-based number. An obvious approach is to evaluate the regression equation for each x -value in the vehicle population, weight the corresponding y -values according to number of vehicles of each species sold, and sum to obtain an estimated sales-weighted average on-road number. For example, suppose there are N vehicle species having respective EPA-based numbers $x_1, x_2, \dots, x_N$ and corresponding sales weights $f_1, f_2, \dots, f_N$. Evaluating the regression equation for x_1 yields an estimated in-use number y_1 , and similarly $x_2, x_3, \dots, x_N$ have corresponds $y_2, y_3, \dots, y_N$. Then the sales-weighted average in-use number can be computed as

$$\bar{y}_{sw} = f_1y_1 + f_2y_2 + \dots + f_Ny_N = \sum_{i=1}^N f_iy_i$$

where

$$\sum_{i=1}^N f_i = 1$$

Substitution for y_i according to

$$y_i = b_0 + b_1x_i$$

gives

$$\begin{aligned}\bar{y}_{sw} &= \sum_{i=1}^N f_i (b_0 + b_1 x_i) \\ &= b_0 \sum_{i=1}^N f_i + b_1 \sum_{i=1}^N f_i x_i \\ &= b_0 + b_1 \bar{x}_{sw}\end{aligned}$$

where

$$\bar{x}_{sw} = \sum_{i=1}^N f_i x_i$$

is by definition the sales-weighted EPA number. Thus, it is seen that one can obtain an estimate of the sales-weighted in-use number merely by evaluating the regression equation for the sales-weighted EPA number. In short, "the average in-use number for all cars is the in-use number for the average car."

Recall, now, that the variance function which expresses the variance of $\hat{y}$ as an estimate of x is of the form of a parabola. Accordingly, intuition would suggest that vehicles having x -values far from the mean of the sample points would have relatively high uncertainty with regard to their estimated in-use numbers and that this fact would tend to inflate the variance of the sales-weighted average fuel economy (fuel consumption). The first part of this supposition is correct: extreme values of x do tend to have a high degree of uncertainty associated with their corresponding $\hat{y}$ estimates. The second part of the supposition is not necessarily true. The variance of a weighted combination of y -value estimates is the same as the variance of the y -value corresponding to the weighted combination of the x -values. This conclusion is derived formally in Appendix II.

What do these findings imply for the estimation of in-use average fuel economy (fuel consumption) or the sales-weighted average shortfall? Most important, perhaps, is the implication that the distribution of vehicle species which make up the data base is not critical, so long as the average EPA-based number for the sample is "near" the sales-weighted

average of the vehicle population. Pushed to its logical conclusion, this assertion is equivalent to saying that there is no need to sample either very high-mpg or very low-mpg vehicles unless there is interest in defining the shortfall as a function of EPA-based fuel-economy.

Quite clearly the above implications are not free of engineering caveats. The conclusions drawn above are based on a slavish belief in the validity of the regression equation and that all vehicles in the sample obey the equation equally well. If differences among vehicle manufacturers or among data sources suggest the need for more than one regression relation, then modifications in the statistical approach would be required. Moreover, bias can occur if the regression model chosen, whether in mpg or gpm space, is not actually linear and consequently does not provide an accurate basis for interpolation.

Finally, there are some engineering and philosophical difficulties in accepting the belief that "the average shortfall for all cars is the same as the shortfall for the average car," especially if the shortfall is to be tracked over time. Technological and design changes are the bases by which fuel economy is improved, and there is no assurance that these changes are the same across the entire mpg spectrum. It seems essential, therefore, that the vehicles included in the data base should provide a realistic sampling of technology even though statistical requirements (under the given assumptions) do not seem to demand it. In reality, there is no incompatibility in the two views if it is realized that different regression equations may result if certain technological innovations are included or excluded from the data base.

6. Summary and Conclusions

Considered in this paper are some of the factors which influence the precision and accuracy with which in-use fuel economy (fuel consumption) can be estimated by regression analysis. Weighting factors peculiar to the vehicle population, the available data base, and the needs and requirements of the analyst are put in perspective.

The limitations of the data base may be such that not all of the requirements of the analyst can be met. It is at this point that decisions must be made with regard to the resolution of conflicting requirements, and these decisions must be reflected in the approach to statistical analysis.

Some of the specific findings pertinent to the assessment of fuel-economy shortfall are itemized below:

- The assessment of fuel-economy shortfall for specific vehicle species imposes different statistical requirements than does the assessment of sales-weighted average shortfall.
- Given a valid regression model, the sales-weighted average in-use fuel economy can be obtained by evaluating the regression equation for the sales-weighted average EPA number, with no loss in precision.
- The regression approach implies that, so far as estimation of average shortfall is concerned, the selection of vehicles to be included in the data base is not critical so long as the average EPA rating for these vehicles is in the vicinity of the corresponding sales-weighted average for the population.
- The regression model implies that for evaluation of shortfall as a function of EPA-based fuel economy it is preferable to concentrate the vehicles in the data base at the extreme ends of the fuel-economy range of interest.

- Data-base restrictions may preclude the realization of best linear unbiased estimation, in which case compromises in the statistical approach must be made in accordance with the practical ends to be served by the analysis.
- Because of the strong dependence of conclusions on the nature of the regression model, it is imperative that the choice of model be in accordance with the best engineering judgment with regard to the relation between in-use and EPA-based fuel economy (fuel consumption).

It is realized that the above conclusions are based on the assumption that the data base is homogeneous, even though it might consist of several data sources. Criteria for assessment of the homogeneity of sources, as well as for assessment of outliers within sources, constitutes a problem to be addressed in another Interpretation Paper.

APPENDIX I

The question of estimating a weighted average of two or more distributions by means of regression analysis has an archetype in the simpler problem discussed herein. That problem concerns estimating a weighted average of the means of two distributions from samples taken randomly from those distributions.

Let the two distributions, referred to hereinafter as the first and second distributions, have means m_1 and m_2 , respectively, and common variance σ^2 . Let N be the total number of observations to be shared between the two samples, with n allocated to the first distribution and $N-n$ allocated to the second distribution. Let $\bar{x}_1$ and $\bar{x}_2$ be the respective sample means.

It is desired to obtain an estimate of the weighted combination

$$m = wm_1 + (1-w)m_2 \quad (I-1)$$

where w is the weight assigned to the first distribution and $1-w$ is the weight assigned to the second distribution. The problem under consideration is to determine the optimum division of N into two parts n and $N-n$ so as to provide a minimum variance estimation of m .

As an estimator of m , consider

$$\bar{x} = w\bar{x}_1 + (1-w)\bar{x}_2 \quad (I-2)$$

Then

$$\sigma_{\bar{x}}^2 = w^2\sigma_{\bar{x}_1}^2 + (1-w)^2\sigma_{\bar{x}_2}^2 \quad (I-3)$$

$$\text{where } \sigma_{\bar{x}_1}^2 = \frac{\sigma^2}{n} = \frac{1}{N} \frac{\sigma^2}{(n/N)} \quad (I-4)$$

$$\text{and } \sigma_{\bar{x}_2}^2 = \frac{\sigma^2}{N-n} = \frac{1}{N} \frac{\sigma^2}{(1-\frac{n}{N})} \quad (I-5)$$

Writing n/N as r , one has from (I-3)

$$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{N} \left(\frac{w^2}{r} + \frac{(1-w)^2}{1-r} \right) \equiv f(r)$$

Minimizing by differentiation of $f(r)$, one has

$$\frac{df(r)}{dr} = \frac{\sigma^2}{N} \left[\frac{-w^2}{r^2} + \frac{(1-w)^2}{(1-r)^2} \right] = 0$$

$$\frac{w^2}{r^2} = \frac{(1-w)^2}{(1-r)^2}$$

$$\frac{w}{r} = \frac{1-w}{1-r}$$

from which

$$r = w$$

and

$$1-r = 1-w$$

In other words, minimum-variance estimation is achieved by allocating the N observations in accordance with the weighting factors w and $1-w$ in (I-1). For example if $w=0.3$, $1-w = 0.7$, and $N=100$, then 30 observations should be allocated to the first distribution and 70 observations to the second.

By an extension of this argument it can be shown that if more than two distributions are involved, observations should be allocated to the various distributions in the same ratio as the weights to be employed in the weighted average. In short, to estimate the sales-weighted on-road fuel economy one should employ a sales-weighted sample.

The importance of the assumption of common variance σ^2 for the several populations needs to be emphasized. A similar analysis can be performed in the case of unequal variances, but it is to be understood that this inequality would affect the resulting optimum weighting factors.

APPENDIX II

We wish to consider here, in the context of regression analysis, the estimation of expected values and variances for linear combinations of random variables. First, however, we require some preliminaries.

1. Some Preliminaries

Let $a_1, a_2, \dots, a_N$ be real constants and let $X_1, X_2, \dots, X_N$ be random variables. A linear combination of these random variables is defined to be

$$Z = a_1X_1 + a_2X_2 + \dots + a_NX_N \quad (\text{II-1})$$

The expected value of Z is

$$\begin{aligned} E[Z] &= E[a_1X_1 + a_2X_2 + \dots + a_NX_N] \\ &= E[a_1X_1] + E[a_2X_2] + \dots + E[a_NX_N] \\ &= a_1 E[X_1] + a_2 E[X_2] + \dots + a_N E[X_N] \end{aligned} \quad (\text{II-2})$$

Suppose $E[X_1] = m_1, E[X_2] = m_2, \dots, E[X_N] = m_N$. Then

$$E[Z] = a_1m_1 + a_2m_2 + \dots + a_Nm_N. \quad (\text{II-3})$$

Thus the expected value of a weighted sum of random variables is the same weighted sum of their expected values.

The variance of the combined random variable Z is also of interest and can be derived from the definition of variance as

$$\text{Var}(Z) = E[Z - E(Z)]^2 = E[Z^2] - [E(Z)]^2 \quad (\text{II-4})$$

Let $N = 2$ in (II-1). Substituting for Z its counterpart in terms of X_1 and X_2 , we have

$$\begin{aligned}
 \text{Var } [Z] &= E \left\{ [a_1 X_1 + a_2 X_2]^2 \right\} - [E(a_1 X_1 + a_2 X_2)]^2 \\
 &= E [a_1^2 X_1^2 + a_2^2 X_2^2 + 2a_1 a_2 X_1 X_2] - [a_1 E(X_1) + a_2 E(X_2)]^2 \\
 &= a_1^2 E [X_1^2] + a_2^2 E [X_2^2] + 2a_1 a_2 E(X_1 X_2) \\
 &\quad - a_1^2 [E(X_1)]^2 - a_2^2 [E(X_2)]^2 - 2a_1 a_2 [E(X_1) E(X_2)] \\
 &= a_1^2 \left\{ E [X_1^2] - [E(X_1)]^2 \right\} + a_2^2 \left\{ E [X_2^2] - [E(X_2)]^2 \right\} \\
 &\quad + 2a_1 a_2 \left\{ E(X_1 X_2) - E(X_1) E(X_2) \right\} \\
 &= a_1^2 \text{Var } X_1 + a_2^2 \text{Var } X_2 + 2a_1 a_2 \text{Cov}(X_1, X_2)
 \end{aligned}$$

More generally, if one has N random variables $X_1, X_2, \dots, X_N$, then

$$\text{Var } [Z] = \sum_{i=1}^N a_i^2 \text{Var } X_i + 2 \sum_{i=1}^{N-1} \sum_{j=i+1}^N a_i a_j \text{Cov } (X_i, X_j) \quad (\text{II-5})$$

Note that if $\text{Cov } (X_i, X_j) = 0$ for all $i \neq j$,

$$\text{Var } [Z] = \sum_{i=1}^N a_i^2 \text{Var } X_i.$$

Thus the variance of a weighted sum of uncorrelated random variables is equal to a weighted sum of their variances, where the weights a_i^2 are the squares of the weights a_i in (II-1). If, in addition,

$$\text{Var } [X_1] = \text{Var } [X_2] = \dots = \text{Var } [X_N] = \sigma^2$$

then

$$\text{Var } [Z] = \sigma^2 \sum_{i=1}^N a_i^2.$$

2. Simple Linear Regression Revisited

The basic least-squares approach to simple linear regression has previously been presented in Application Paper No. 2.* There it was shown that the variances of the intercept b_0 and slope b_1 in the linear equation

$$z(x) = b_0 + b_1 x \quad (\text{II-6})$$

can be expressed in terms of the inverse of the matrix $X'X$ of the least-squares normal equations, as defined in the cited Application Paper. Thus $\text{Var}(\underline{b})$, subsequently called $v(\underline{b})$, is given by

$$\text{Var}(\underline{b}) = [X'X]^{-1} \sigma^2 \quad (\text{II-7})$$

where

$$\underline{b} = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix}$$

is the vector of coefficients of the model and σ^2 is the error variance (which usually must be estimated from the residuals around the regression line).

Consider an arbitrary point x along the axis of the independent variable and define a vector $\underline{x}$ as

$$\underline{x} = \begin{bmatrix} 1 \\ x \end{bmatrix}.$$

The estimated value of z at that point is

$$\hat{z} = \underline{x}' \underline{b} = \begin{bmatrix} 1 & x \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} = b_0 + b_1 x \quad (\text{II-8})$$

as in (II-6).

* H. T. McAdams, Fuel Economy of Automotive Vehicles: A Compendium of Statistical Applications, Falcon Research & Development Company, Buffalo, NY (June 1978).

The random variable $\hat{z}$ has expected value $\underline{x}'\underline{\beta}$, where $\underline{\beta}$ is the vector of (unknown) coefficients of the model. The variance of $\hat{z}$ is, by definition,

$$\begin{aligned}\text{Var}(\hat{z}) &= E [\hat{z} - E(\hat{z})] [\hat{z} - E(\hat{z})]' \\ &= E [\underline{x}'\underline{b} - \underline{x}'\underline{\beta}] [\underline{x}'\underline{b} - \underline{x}'\underline{\beta}]' \\ &= E [\underline{x}'(\underline{b} - \underline{\beta})] [\underline{x}'(\underline{b} - \underline{\beta})]' \\ &= E [\underline{x}'(\underline{b} - \underline{\beta})(\underline{b} - \underline{\beta})' \underline{x}] \\ &= \underline{x}' E [(\underline{b} - \underline{\beta})(\underline{b} - \underline{\beta})'] \underline{x}\end{aligned}$$

But $E(\underline{b} - \underline{\beta})(\underline{b} - \underline{\beta})' = V(\underline{b})$. Therefore,

$$\text{Var}(\hat{z}) = \underline{x}'(X'X)^{-1} \underline{x}\sigma^2 \quad (\text{II-9})$$

The result (II-9) can be made more explicit by noting that, in simple linear regression,

$$X'X = \begin{bmatrix} N & \sum_{i=1}^N x_i \\ \sum_{i=1}^N x_i & \sum_{i=1}^N x_i^2 \end{bmatrix} \quad (\text{II-10})$$

and (dropping the subscripts for convenience),

$$(X'X)^{-1} = \frac{1}{N \sum x^2 - (\sum x)^2} \begin{bmatrix} \sum x^2 & -\sum x \\ -\sum x & N \end{bmatrix} \quad (\text{II-11})$$

In the interest of simplifying the inverse matrix, it is convenient to replace each x_i in (II-10) by $x_i - \bar{x}$, where $\bar{x}$ is the mean of the x_i . Without any loss of generality one can then write

$$X'X = \begin{bmatrix} N & \sum(x - \bar{x}) \\ \sum(x - \bar{x}) & \sum(x - \bar{x})^2 \end{bmatrix} = \begin{bmatrix} N & 0 \\ 0 & \sum(x - \bar{x})^2 \end{bmatrix} \quad (\text{II-12})$$

Then

$$(X'X)^{-1} = \frac{1}{N \Sigma(x-\bar{x})^2} \begin{bmatrix} \Sigma(x-\bar{x})^2 & 0 \\ 0 & N \end{bmatrix}$$

and by (II-7),

$$\text{Var } (b_0) = \frac{\sigma^2}{N} \quad (\text{II-13})$$

and

$$\text{Var } (b_1) = \frac{\sigma^2}{\Sigma(x-\bar{x})^2} \quad (\text{II-14})$$

Also, by virtue of (II-9),

$$\begin{aligned} \text{Var } [\hat{z}(x)] &= \begin{bmatrix} 1 & (x-\bar{x}) \end{bmatrix} \begin{bmatrix} 1/N & 0 \\ 0 & \Sigma(x-\bar{x})^2 \end{bmatrix} \begin{bmatrix} 1 \\ (x-\bar{x}) \end{bmatrix} \sigma^2 \\ &= \left\{ \frac{1}{N} + \frac{(x-\bar{x})^2}{\Sigma(x-\bar{x})^2} \right\} \sigma^2 \end{aligned} \quad (\text{II-15})$$

where $\hat{z}(x)$ is the estimate of z for specified x as in (II-6).

Requisite to later considerations is an estimate of the covariance between $\hat{z}_1 = \hat{z}(x_1)$ and $\hat{z}_2 = \hat{z}(x_2)$ --that is, the covariance of estimated values of z at two locations x_1 and x_2 along the x -axis. We proceed in a manner parallel to the development preceding (II-9).

By definition,

$$\begin{aligned} \text{Cov}(\hat{z}_1, \hat{z}_2) &= E [\hat{z}_1 - E(\hat{z})] [\hat{z}_2 - E(\hat{z}_2)] \\ &= E [\underline{x}_1' \underline{b} - \underline{x}_1' \underline{\beta}] [\underline{x}_2' \underline{b} - \underline{x}_2' \underline{\beta}]' \\ &= E [\underline{x}_1' (\underline{b} - \underline{\beta})] [\underline{x}_2' (\underline{b} - \underline{\beta})]' \\ &= E [\underline{x}_1' (\underline{b} - \underline{\beta}) (\underline{b} - \underline{\beta})' \underline{x}_2] \\ &= \underline{x}_1' E [(\underline{b} - \underline{\beta}) (\underline{b} - \underline{\beta})] \underline{x}_2 \\ &= \underline{x}_1' V(\underline{b}) \underline{x}_2 \end{aligned}$$

or

$$\text{Cov}(\hat{z}_1, \hat{z}_2) = \underline{x}_1' (X'X)^{-1} \underline{x}_2 \sigma^2 \quad (\text{II-16})$$

Invoking, now, the simplifications of (II-11), one has, analogous to (II-15),

$$\begin{aligned} \text{Cov}(\hat{z}_1, \hat{z}_2) &= \begin{bmatrix} 1 & (x_1 - \bar{x}) \end{bmatrix} \begin{bmatrix} 1/N & 0 \\ 0 & \frac{1}{\sum (x - \bar{x})^2} \end{bmatrix} \begin{bmatrix} 1 \\ (x_2 - \bar{x}) \end{bmatrix} \sigma^2 \\ &= \left\{ \frac{1}{N} + \frac{(x_1 - \bar{x})(x_2 - \bar{x})}{\sum (x - \bar{x})^2} \right\} \sigma^2 \end{aligned} \quad (\text{II-17})$$

3. Variance of a Weighted Sum of $\hat{z}_1$ and $\hat{z}_2$

With the preliminaries of the previous two sections one can now consider the variance of the weighted sum

$$\hat{z} = w \hat{z}_1 + (1-w) \hat{z}_2 \quad (\text{II-18})$$

where z_1 and z_2 are estimates of the dependent variable obtained at two values x_1 and x_2 of the independent variable. In accordance with (II-5),

$$\begin{aligned} \text{Var}[\hat{z}] &= w^2 \text{Var}[\hat{z}_1] + (1-w)^2 \text{Var}[\hat{z}_2] \\ &\quad + 2w(1-w) \text{Cov}[\hat{z}_1, \hat{z}_2] \end{aligned} \quad (\text{II-19})$$

Substitution of the results of (II-15) and (II-17) into (II-19) gives:

$$\begin{aligned} \frac{\text{Var } [\hat{z}]}{\sigma^2} &= w^2 \left\{ \frac{1}{N} + \frac{(x_1 - \bar{x})^2}{\Sigma(x - \bar{x})^2} \right\} + (1-w)^2 \left\{ \frac{1}{N} + \frac{(x_2 - \bar{x})^2}{\Sigma(x - \bar{x})^2} \right\} \\ &\quad + 2w(1-w) \left\{ \frac{1}{N} + \frac{(x_1 - \bar{x})(x_2 - \bar{x})}{\Sigma(x - \bar{x})^2} \right\} \quad (\text{II-20}) \end{aligned}$$

Considering first the terms in $1/N$,

$$1/N \left\{ w^2 + (1-w)^2 + 2w(1-w) \right\} = 1/N \quad (\text{II-21})$$

Next, consider the terms in $\frac{1}{\Sigma(x - \bar{x})^2}$:

$$\begin{aligned} \frac{1}{\Sigma(x - \bar{x})^2} &\left\{ w^2(x_1 - \bar{x})^2 + (1-w)^2(x_2 - \bar{x})^2 + 2w(1-w)(x_1 - \bar{x})(x_2 - \bar{x}) \right\} \\ &= \frac{[w(x_1 - \bar{x}) + (1-w)(x_2 - \bar{x})]^2}{\Sigma(x - \bar{x})^2} = \frac{[w_1 x_1 + (1-w) x_2 - \bar{x}]^2}{\Sigma(x - \bar{x})^2} \quad (\text{II-22}) \end{aligned}$$

Combining (II-21) and (II-22), one has

$$\begin{aligned} \text{Var } [\hat{z}] &= \text{Var } [w \hat{z}_1 + (1-w) \hat{z}_2] \\ &= \left\{ \frac{1}{N} + \frac{[w_1 x_1 + (1-w) x_2 - \bar{x}]^2}{\Sigma(x - \bar{x})^2} \right\} \sigma^2 \quad (\text{II-23}) \end{aligned}$$

In conclusion, (II-23) can be interpreted as follows. As the problem was initially stated, equation (II-6) was evaluated at x_1 and x_2 to obtain the estimates $\hat{z}_1$ and $\hat{z}_2$. These were then weighted by w and $1-w$, respectively, and the variance of this weighted combination was derived as (II-23). Note, however, that one could take a weighted average of x_1 and x_2 and evaluate (II-6) for this weighted value

$$w x_1 + (1-w) x_2.$$

The resulting estimate of $\hat{z}$ would be the same, because

$$\begin{aligned} w_1 \hat{z}_1 + (1-w) \hat{z}_2 &= w (b_0 + b_1 x_1) + (1-w) (b_0 + b_1 x_2) \\ &= b_0 + b_1 [w x_1 + (1-w) x_2] \end{aligned}$$

Note that if one were to substitute $w x_1 + (1-w) x_2$ for x in (II-15), the result is directly (II-23). Thus the variance of the weighted sum of $\hat{z}_1$ and $\hat{z}_2$ is the same as the variance of $\hat{z}$ corresponding to the weighted sum of x_1 and x_2 .

By an extension of the above argument it can be shown that a similar result holds for a weighted average of any number k of estimates $\hat{z}_i$ ($k = 1, 2, \dots, k$).

Whittaker

Interpretation Paper No. 3
Project No. 8431
U.S. Dept. of Energy
October 15, 1979
H. T. McAdams

WEIGHTING FACTORS REVISITED: HOW THEY ARE INFLUENCED
BY ASSUMPTIONS MADE IN THE REGRESSION MODEL

1. Introduction

Regression analysis has come to be the primary technique employed in the assessment of the so-called fuel-economy "shortfall" which characterizes the discrepancy between EPA fuel-economy values for automotive vehicles and actual, in-use experience for those vehicles. Its use is based on the belief that some relation, albeit an unknown one, exists between the two sets of numbers and that this relationship is, in some sense, consistent over the vehicle population.

To implement the regression approach one must make certain assumptions with regard to the form of the regression equation, the nature of the error distribution, and the basis of the estimation procedure employed in determining the regression-equation parameters. Though these aspects of regression analysis have been considered in an earlier Interpretation Paper¹, it is of interest to explore alternative assumptions and how they impact upon conclusions drawn by way of the regression approach.

¹ H. T. McAdams, "Weighting-Factor Considerations in the Regression Analysis of Vehicle Fuel Usage," Interpretation Paper No. 2 (August 1, 1979).

2. Regression Model Alternatives

As typically employed in data analysis, regression techniques fall within the purview of the theory of linear statistical models. The term linear requires clarification, however.

Let it be assumed that two mathematical variables, x and y , are related as follows:

$$y = \beta_1 f_1(x) + \beta_2 f_2(x) + \dots + \beta_r f_r(x) \quad (1)$$

where $f_1, f_2, \dots, f_r$ are of arbitrary form but are linearly independent. Equation (1) is said to be linear in $\beta_1, \beta_2, \dots, \beta_r$ even though the functions $f_1, f_2, \dots, f_r$ are nonlinear in the variable x . Thus (1) can be regarded as a linear model in which the functions $f_1, f_2, \dots, f_r$ are regarded as a basis capable of spanning a certain class of functional relations between the variables y and x .

In regression analysis the parameters $\beta_1, \beta_2, \dots, \beta_r$ are unknown and must be estimated from data which, in the present context, are derived from EPA tests of fuel economy and in-use fuel requirements.

Let x represent fuel economy² as determined by the EPA certification test, and let y denote in-use fuel economy. Ordered pairs (x_i, y_i) of numbers are generated if one associates with the i th vehicle an EPA number x_i and an in-use number y_i . Ideally, the set of ordered pairs obtained by this means constitutes a function which maps the EPA value, as preimage, into an in-use number, as image. In reality, however, the relation between y and x is a many-to-one mapping. For a given certification value x_i , an entire distribution of in-use values arises. Each vehicle belonging to a certification class of vehicles has its own peculiarities, and is most likely driven at different speeds, over different terrain, and under a host of other conditions which are different from those to which all other vehicles of the class are subjected. A reasonable approach is to regard these peculiarities as random variables hopelessly beyond control and to abstract from this array of many-to-one correspondences an idealized relation which minimizes, in some sense, the uncertainty inherent in the prediction of y , given a measured value of x . Typically this minimization is achieved by least squares.

² Fuel economy is used here only for purposes of exemplifying methodology. Similar considerations apply if one chooses to perform the analysis in fuel-consumption space.

2.1 The Simple Linear Regression Model

A formal representation of the aforementioned many-to-one mapping is afforded by the linear statistical model

$$y_i = \beta_0 + \beta_1 x_i + e_i \quad (2)$$

where e_i is the outcome of an error random variable ϵ_i having the properties

$$\left. \begin{aligned} E(\epsilon_i) &= 0 \\ E(\epsilon_i^2) &= \sigma^2 \\ E(\epsilon_i \epsilon_j) &= 0, i \neq j \end{aligned} \right\} \quad (3)$$

where E denotes expectation. The assumption that $E(\epsilon_i^2) = \sigma^2$ is equivalent to the assumption of homoscedasticity--that is, that the error variance is constant for all values of x . It is this model which is usually assumed to underlie simple linear regression. In the present application x_i is a fixed value stemming from the EPA test but, by virtue of the random variable ϵ_i , the value of y_i is the outcome of a random variable.

Given the assumptions of (2), one can arrive at certain consequences pertaining to statistical inference, such as the precision with which in-use fuel economy can be estimated, given a specified value of fuel economy as determined by the EPA test. A related question pertains to estimation of aggregate, in-use fuel economy given--say--the sales-weighted average fuel economy as determined by EPA test. Of particular concern in this connection is how best to allocate vehicles in the sample when only a limited number of vehicles can be tested in use. Should the vehicles be selected so as to span the entire range of EPA certification values, or should they be selected so as to cluster in the vicinity of the sales-weighted average?

These questions have been previously considered in Interpretation Paper No. 2 (op cit), and it will be informative to recall some of the findings of that paper. If our concern is primarily with the estimation of in-use average fuel economy, the distribution of vehicle species which make up the data base is not critical, so long as the average

EPA-based number is "near" the sales-weighted average of the vehicle population. Thus there would seem to be no need to sample either very high-mpg or very low-mpg vehicles unless there is interest in defining in-use fuel economy as a function of EPA-based fuel-economy. This conclusion is based on the fact that, under the assumptions of the model, "the average fuel economy for all cars is the same as the fuel economy for the average car."

Statistical considerations notwithstanding, "common sense" tends to militate against minimizing the importance of a rational choice of vehicle species to be included in the data base. It is the purpose of this section to provide an explicit basis for these reservations in the form of an error model believed to be more realistic than the one assumed above.

To arrive at such an error model let us examine carefully the implications of equation (2). Note that the regression model as stated provides no means of incorporating information pertaining to any engineering or technological distinctions between vehicles species exhibiting the same EPA-based fuel economy. It is tacit in the model that, except for random error contributions, the only factor which influences in-use fuel economy is the EPA certification fuel-economy value. A corollary implication is that all vehicle species exhibiting a given EPA-based fuel economy number are equally "representative" for that number and respond in the same way to those factors which give rise to fuel-economy shortfall.

The limitations of this view are dramatized by comparing line entries in the Gas Mileage Guide. For example, based on the 1978 Gas Mileage Guide, Table I lists a variety of vehicles all rated at 20 mpg combined EPA fuel economy. Note that these vehicles range from mini-compact two-door hatchbacks to four-door midsize wagons and from 171 CID 6-cylinder to 301 CID 8-cylinder engines. They include both automatic and manual transmissions and 1, 2, and 4-barrel carburetors. In the table one might also have included pick-up trucks and vans rated at 20 mpg.

Table I
VEHICLES RATED AT 20 MPG COMBINED EPA
FUEL ECONOMY (1978 GAS MILEAGE GUIDE)

<u>Size Class</u>	<u>Manufacturer Car Line</u>	<u>Engine CID/Cyl.</u>	<u>A/M*</u>	<u>Fuel System</u>	<u>Body Type Int./Trunk(ft.³)</u>
Mini-Compact	Ford Pinto	171/6	A	2	Htbk - 74/9
Sub-Compact	Chevrolet Monza	305/8	A	2	Htbk - 78/10
Compact	Chevrolet Nova	250/6	A	1	4 Dr - 96/13
Midsize	Dodge Monaco	225/6	A	1	2 Dr - 95/15
Midsize	Pontiac Grand Prix	301/8	A	4	2 Dr - 96/16
Small Wagon	Mercury Bobcat	171/6	A	2	2 Dr - 78/31
Midsize Wagon	Chrysler LeBaron	225/6	M	2	4 Dr - 98/39
Midsize Wagon	Ford Fairmont	200/6	A	1	4 Dr - 98/43

* Transmission (A) Automatic, (M) Manual

It is not difficult to appreciate the fact that the vehicles listed in Table I, though found equivalent in the EPA test, might be expected to exhibit different in-use fuel economy performance. In addition to engineering distinctions among the vehicles, their use patterns would ostensibly depend on the class of consumer associated with the various vehicles, and in-use fuel economy would reflect these differences.^{3, 4} For example, purchasers of 4-door vehicles are usually older than purchasers of 2-door vehicles, and the use pattern for "the family station wagon" is different from that of--say--a small hatchback or a midsize sedan. In short, it is postulated that each species of vehicle in Table I has a species-specific "error" relative to the EPA fuel-economy number and that this error is determined by the engineering and use-environment factors peculiar to that species. This error is in addition to the random errors which arise as the result of sampling variation associated with individual vehicles within a vehicle species.

2.2 Alternative Error Models

In view of the above considerations, it is reasonable to propose an alternative error model for equation (2). For a particular EPA number x_i , the associated design and use-environment variables may conspire to place the corresponding average, in-use fuel economy either above or below the trend line. Thus, instead of the error properties (3), one might assume:

$$\left. \begin{aligned} E(\epsilon_i) &= \alpha_i \\ E(\epsilon_i - \alpha_i) &= \sigma^2 \\ E(\epsilon_i - \alpha_i)(\epsilon_j - \alpha_j) &= 0, i \neq j \end{aligned} \right\} \quad (4)$$

Under these assumptions there remains a certain lack of fit even if it were possible to eliminate completely the random error contribution. This lack of fit results from fixed errors associated with specific vehicles and their related use environments.

³ J. D. Murrell and K. Hellman, "In-Use and EPA MPG: Factors Influencing the Comparison," SAE Paper 790934 (1979)

⁴ M. F. Hunt, "Impact of Fuel Economy Information on Automobile Purchase Choice," SAE Paper 790933 (1979)

A further elaboration of this error concept becomes apparent if one considers the various sources of data which comprise the in-use, fuel economy data base. Each source of data has an associated mix of vehicle species and use environments. Consequently, each source can be regarded as bringing to the regression analysis its own fixed error contribution.

3. Estimation Implications of Alternative Error Models

The implications of estimation theory for fuel-economy assessment have been considered in Interpretation Paper No. 2 (op. cit.) but only under the error assumptions given by (3). The alternative assumptions given by (4) are equivalent to the inclusion of both fixed and random effects in the model and thus give rise to what is often referred to as a "mixed" model in standard statistical texts.⁵ The implications of the mixed model assumption are discussed below.

As a point of departure for this discussion, it is informative to recall the regression model (2), which can be written in continuous form as

$$Y = \beta_0 + \beta_1 x + \epsilon \quad (5)$$

where ϵ and Y are random variables. Under the error assumption (3), the conditional expected value of Y for a given value of x is

$$\left. \begin{aligned} E[Y|x] &= E[(\beta_0 + \beta_1 x + \epsilon)|x] \\ &= \beta_0 + \beta_1 x + E[\epsilon|x] \\ &= \beta_0 + \beta_1 x \end{aligned} \right\} \quad (6)$$

since it is assumed that $E[\epsilon|x] = 0$ for all x . Thus (2) together with (3) provides unbiased estimation of y for a given value of x .

Let us now consider (2) with the error assumptions of (4). Then

$$\begin{aligned} E[Y|x] &= E[(\beta_0 + \beta_1 x + \epsilon)|x] \\ &= \beta_0 + \beta_1 x + E[\epsilon|x]. \end{aligned}$$

⁵ See, for example, G. W. Snedecor, Statistical Methods, Section 10.15, "Samples Within Samples. Mixed Model," The Iowa State College Press, Ames, Iowa (1956).

But

$$E \left[\epsilon | x \right] = \alpha(x), \text{ and}$$

$$E \left[Y | x \right] = \beta_0 + \beta_1 x + \alpha(x), \quad (7)$$

where $\alpha(x)$ denotes a fixed error associated with a specific value of x and, ipso facto, with a specific vehicle species or combination of vehicle species having the same EPA rating.* Thus, $y = \beta_0 + \beta_1 x$ does not provide an unbiased estimate of y for a given x . The residual, $\alpha(x)$, is a departure from linearity and thus connotes a lack-of-fit term which can affect the estimation of β_0 and β_1 to a degree which depends on the selection of the x_i included in the data base.

* In the event several species are involved, the variance σ^2 will reflect differences among species as well as sampling variation within species.

4. Weighting Factor Considerations

To investigate how the choice of vehicles affects the estimation process, it is convenient to view the collection of in-use vehicle species as a finite population, where the term "species" can be thought of as a line entry in the Gas Mileage Guide. The number of vehicles sold in a given species, expressed as a fraction of the total vehicles sold, gives the sales weight for that species. If a vehicle is selected randomly from the population, this sales weight would be the probability that the vehicle belongs to the species in question.

Suppose there are K vehicle species in the population with sales weights $f_1, f_2, \dots, f_K$. Within each species individual vehicles display an associated fuel economy, but one can combine the expected values or means for each species with the corresponding sales weights to obtain a sales-weighted average fuel economy. Let $m_1, m_2, \dots, m_K$ denote the true in-use species means. Note that these means can be computed only if the fuel economy for every vehicle in the population is known. Nevertheless, one can write conceptually

$$m = f_1 m_1 + f_2 m_2 + \dots + f_K m_K = \sum_{i=1}^K f_i m_i \quad (8)$$

for the population, sales-weighted fuel economy. In reality, of course, the m_i are not known but must be estimated from a sample much smaller than the total number of vehicles in the population. It may, indeed, turn out that not all of the species will be represented in the sample, and this fact could introduce bias into the estimation of the sales-weighted fuel economy.

In the remainder of this section attention will be addressed to the estimation of sales-weighted average fuel economy under two basic assumptions. In the first, it will be assumed that every species of vehicle present in the population will be represented, to some degree at least, in the sample. In the second, it will be assumed that only a subset of all vehicle species are present in the sample and that some interpolation procedure, such as regression analysis, must be invoked to fill the data gaps. These two alternatives will be referred to as complete and incomplete sampling, respectively.

4.1 Complete Sampling

Let us consider, first, the ideal circumstances under which estimation of sales-weighted, in-use fuel economy could occur. Let $N_1, N_2, \dots, N_K$ denote the numbers of vehicles in the population for each of the K species, and let $n_1, n_2, \dots, n_K$ denote the corresponding number of vehicles in the sample. Define

$$N_1 + N_2 + \dots + N_K = \sum_{i=1}^K N_i = N$$

as the total number of vehicles in the population and

$$n_1 + n_2 + \dots + n_K = \sum_{i=1}^K n_i = n$$

as the total number of vehicles in the sample. Then $f_i = N_i/N$ denotes the sales weight for the i th vehicle species and $g_i = n_i/n$ denotes the corresponding sample weight. Letting $\bar{y}_1, \bar{y}_2, \dots, \bar{y}_K$ denote the fuel-economy sample means for the respective species, one wishes to examine the conditions under which the sample means can be combined to obtain an unbiased estimate of the population, sales-weighted average fuel economy. Note that the sample means need not be combined computationally in accordance with their weights on the basis of relative number of vehicles in the sample. A second question concerns the allocation of vehicles in the sample in such a way that the estimated sales-weighted average fuel economy has minimum variance.

4.1.1 The Question of Unbiasedness

Let us first examine the question of unbiasedness. Let $\bar{y}_1, \bar{y}_2, \dots, \bar{y}_K$ denote sample means for the K species without regard to the number of vehicles on which these means are based. Under any circumstances it can be assumed that

$$E[\bar{y}_i] = \mu_i \quad i = 1, 2, \dots, K$$

--that is, that the sample means are unbiased estimators of the population means for the several species. Choose arbitrary weighting

factors $h_1, h_2, \dots, h_K$ with $\sum_{i=1}^K h_i = 1$ and write

$$\hat{m} = h_1 \bar{y}_1 + h_2 \bar{y}_2 + \dots + h_K \bar{y}_K = \sum_{i=1}^K h_i \bar{y}_i \quad (9)$$

as an estimator of the true, sales-weighted fuel economy m . Taking expectations, one has

$$\left. \begin{aligned} E[\hat{m}] &= E \left[\sum_{i=1}^K h_i \bar{y}_i \right] \\ &= \sum_{i=1}^K h_i E[\bar{y}_i] \\ &= \sum_{i=1}^K h_i m_i \end{aligned} \right\} \quad (10)$$

in comparison with

$$m = \sum_{i=1}^K f_i m_i \quad (11)$$

For unbiasedness it is necessary that $E[\hat{m}] = m$ or that

$$\sum_{i=1}^K h_i m_i = \sum_{i=1}^K f_i m_i, \quad (12)$$

which is assured, of course, if

$$h_i = f_i \quad \text{for all } i.$$

In other words, the sample means should be combined by assigning them weights equal to the corresponding population sales weights. Note, however, that for a given set of m_i this choice of weights does not preclude other choices satisfying (12). This fact becomes important under circumstances in which it is not possible to sample all vehicle species in the population. It is, of course, this situation which usually applies and which is the reason for the extensive use of regression analysis in the assessment of in-use fuel economy.

4.1.2 The Question of Minimum Variance

Let us now consider the question of obtaining the best linear unbiased estimator (BLUE) for the sales-weighted average fuel economy m , where "best" implies minimum variance. Let $\bar{y}_1, \bar{y}_2, \dots, \bar{y}_K$ denote sample means for the K species, and let $n_1, n_2, \dots, n_K$ be the corresponding sample sizes. Then the variance of $\bar{y}_i$ is given by

$$\text{Var} [\bar{y}_i] = \sigma_i^2 / n_i \quad (13)$$

where σ_i^2 is the population variance for the i th vehicle species. If one assumes homoscedasticity, (13) becomes

$$\text{Var} [\bar{y}_i] = \sigma^2 / n_i \quad (14)$$

where σ^2 is the variance common to all vehicle species.

Under the assumption that the vehicle species are statistically independent, one has from (9)

$$\left. \begin{aligned} \text{Var} [\hat{m}] &= h_1^2 \text{Var} [\bar{y}_1] + h_2^2 \text{Var} [\bar{y}_2] \\ &\quad + \dots + h_K^2 \text{Var} [\bar{y}_K] \\ &= h_1^2 \frac{\sigma_1^2}{n_1} + h_2^2 \frac{\sigma_2^2}{n_2} + \dots + h_K^2 \frac{\sigma_K^2}{n_K} \end{aligned} \right\} \quad (15)$$

or, if one assumes homoscedasticity,

$$\text{Var} [\hat{m}] = \left[\frac{h_1^2}{n_1} + \frac{h_2^2}{n_2} + \dots + \frac{h_K^2}{n_K} \right] \sigma^2 \quad (16)$$

Writing $n_i/n = g_i$, one can write (16) as

$$\text{Var } [\hat{m}] = \left[\frac{h_1^2}{g_1} + \frac{h_2^2}{g_2} + \dots + \frac{h_K^2}{g_K} \right] \frac{\sigma^2}{n}.$$

Also, for unbiasedness, $h_i = f_i$, so that

$$\text{Var } [\hat{m}] = \frac{\sigma^2}{n} \sum_{i=1}^K f_i^2 / g_i \quad (17)$$

Thus, for minimum variance, it is necessary to choose the g_i so as to minimize the quantity

$$Q = \sum_{i=1}^K f_i^2 / g_i \quad (18)$$

where the f_i^2 are fixed constants.

Consider (18) and the fact that $\sum_{i=1}^K g_i = 1$. Then one can write

$$g_K = 1 - \sum_{i=1}^{K-1} g_i \quad (19)$$

and can write (18) as

$$Q = \frac{f_1^2}{g_1} + \frac{f_2^2}{g_2} + \dots + \frac{f_{K-1}^2}{g_{K-1}} + \frac{f_K^2}{1 - \sum_{i=1}^{K-1} g_i} \quad (20)$$

Differentiating Q with respect to g_j ($j = 1, 2, \dots, K-1$) and

equating to zero one obtains

$$\frac{\partial Q}{\partial g_j} = -\frac{f_j^2}{g_j^2} + \frac{f_K^2}{\left(1 - \sum_{i=1}^{K-1} g_i\right)^2} = 0$$

from which

$$\frac{f_j^2}{g_j^2} = \frac{f_K^2}{\left(1 - \sum_{i=1}^{K-1} g_i\right)^2} = \frac{f_K^2}{g_K^2}. \quad (21)$$

Therefore,

$$\frac{f_1}{g_1} = \frac{f_2}{g_2} = \dots = \frac{f_K}{g_K} = c \quad (22)$$

or

$$g_j = c f_j \quad j = 1, 2, \dots, K$$

where c is a proportionality constant. Since $g_j = n_j/n$, one has

$$n_j = c n f_j. \quad (23)$$

In words, the number of vehicles of each species should be proportional to the fraction of the population comprised by that species.

4.2 Incomplete Sampling

In the case of incomplete sampling some of the n_i are zero--that is, there are some vehicle species in the population which are not represented in the sample. It is the intent in this section to address the questions of unbiasedness and sampling efficacy under these incomplete sampling conditions. The discussion will give considerable attention to the effectiveness of linear regression analysis when the linear model is perturbed by lack-of-fit errors attached to individual species of vehicles.

In section 4.1.1 it was noted that if one has species mean estimates $\bar{y}_1, \bar{y}_2, \dots, \bar{y}_K$ and constructs an estimator for the sales-weighted mean as

$$\hat{m} = \sum_{i=1}^K h_i \bar{y}_i,$$

then an unbiased estimator is obtained if $h_i = f_i$, where f_i is the sales fraction for the i th species. A more general relation is provided by (10), however, in which

$$E [\hat{m}] = \sum_{i=1}^K h_i m_i$$

and the fact that one must have

$$E [\hat{m}] = m = \sum_{i=1}^K f_i m_i.$$

For a fixed set of m_i , therefore, there are many choices of h_i for which

$$\sum_{i=1}^K h_i m_i = m$$

but to exploit this notion one would need to know m as well as the m_i . Quite clearly, such information is not generally available, but accumulated experience and engineering judgment can provide a heuristic basis for selecting vehicle species subsets which are fairly "representative" of the set of all vehicle species. A regression equation based on such selection will serve as an unbiased estimator for the sales-weighted, in-use average fuel economy provided the species-specific errors sum to zero.

4.2.1 An Illustrative Example

Though the problem presented above can be addressed formally and analytically, it will be informative to present the issues involved by way of example. For this purpose, we will assume a "mini-population" consisting of only 5 vehicle species. It will be assumed that the samples consist of subsets of 3 species selected from among the 5 in the population. Further, hypothetical "sales weights" will be assumed for the population, and it will also be assumed that the number of vehicles sampled in each species is relatively large. Thus, if one employs the sample means for the species as input data (i.e., if we work at the Level 2 basis of aggregation), the number of vehicles per species will be large enough to eliminate, for practical purposes, the random error contributions ϵ_j . Our intent, of course, is to investigate how the species-related errors, the α_j , affect the regression line and the corresponding estimate of the sales-weighted, average, in-use fuel economy. The computations are performed in mpg space, but it is to be understood that a completely equivalent analysis could be performed in gpm space.

Table 2 presents the hypothetical data on which the example is based.

Table 2
HYPOTHETICAL FUEL ECONOMY DATA

Vehicle Species, i	Sales Fraction, f_i	EPA-Based Fuel Economy (mpg)	In-Use Fuel Economy, m_i (mpg)
1	0.1	15	13
2	0.3	20	19
3	0.2	25	24
4	0.3	30	25
5	0.1	35	29

First, let us compute the actual or true sales-weighted average, in-use mpg. Note that the calculation can be made in this case because it is assumed that all needed inputs are available. Thus, we have, for the EPA ratings,

$$(0.1)(15) + (0.3)(20) + (0.2)(25) + (0.3)(30) + (0.1)(35) = 25.0$$

and, for the in-use ratings,

$$(0.1)(13) + (0.3)(19) + (0.2)(24) + (0.3)(25) + (0.1)(29) = 22.2$$

Thus, 22.2 mpg is the "true," sales-weighted, average, in-use mpg.

Next, let us compute a regression-equation estimate for the sales-weighted, average, in-use mpg. Taking an unweighted approach, one obtains the regression equation

$$y = 3.0 + 0.76x$$

with sales-weighted, in-use, average 22.0 mpg. Note that this number differs from the "theoretical" value of 22.2 mpg. Suppose, however, that one performs a weighted, least-squares regression in which the weighting factors are the sales weights. The resulting regression equation is

$$y = 4.35 + 0.714x$$

with sales-weighted, in-use, average 22.2 mpg in agreement with the theoretical value.

The regression lines as computed by the unweighted and weighted methods are shown in Figure 1. The unweighted and weighted residuals are displayed for the two regressions in Table 3.

Table 3
RESIDUALS FOR HYPOTHETICAL REGRESSIONS

EPA-Rated mpg	Sales Weights	Residuals from Unweighted Equation		Residuals from Weighted Equation	
		Unweighted Residuals	Weighted Residuals	Unweighted Residuals	Weighted Residuals
15	0.1	+ 1.4	+ 0.14	+ 2.06	+ 0.206
20	0.3	- 0.8	- 0.24	- 0.37	- 0.111
25	0.2	- 2.0	- 0.40	- 1.80	- 0.360
30	0.3	+ 0.8	+ 0.24	+ 0.77	+ 0.231
35	0.1	+ 0.6	+ 0.06	+ 0.34	+ 0.034
Sum of Residuals		0.0	- 0.02	+ 1.0	0.0

FIGURE 1 Regression Relations for a Postulated Finite Vehicle Population

Note that, for the unweighted analysis, the weighted sum of residuals is 0.2, the amount by which the estimated sales-weighted, average, in-use mpg is biased. In the weighted analysis, of course, the weighted sum of residuals is zero, and there is no bias.

Suppose, now, that one uses as the sample only the three points corresponding to EPA ratings of 15, 25 and 35 mpg (i.e., species 1, 3, and 5). The unweighted equation is

$$y = 2.0 + 0.8x$$

with estimated sales-weighted, in-use average mpg of 22. A weighted equation can be obtained by applying the population weights appropriate to the points included in the sample. This equation is

$$y = 2.5 + 0.8x$$

with corresponding average of 22.5 mpg. The "true," sales-weighted, mean in-use value is 22.2 mpg. Though the sales-weighted residuals sum to zero for the sample, they do not sum to zero for the population.

Note that, so far as the sample is concerned, the two species not included could have had arbitrary in-use values; accordingly, the success of the sample-based equation depends on the degree to which it approximates the weighted equation based on the entire population. Whether the weighted or unweighted equations perform better, therefore, depends on the particular subset of data chosen.

It is possible to form 10 distinct subsets of 3 from the 5 species included in the hypothetical population. Five of these, selected so as to include extremes, are tabulated in Table 4, together with the estimated, sales-weighted, average, in-use fuel economies as based on the unweighted and weighted approaches.

Table 4
SUBSETS OF HYPOTHETICAL DATA

Subset	Sales-Weighted Average In-Use MPG Based On	
	Unweighted Analysis	Weighted Analysis
1, 2, 3	24.3	24.2
1, 3, 5	22.0	22.5
1, 4, 5	21.0	21.0
2, 3, 4	22.7	22.5
3, 4, 5	23.5	23.5

It is seen that the estimated averages vary from 21.0 mpg to 24.3 mpg, whereas the "true" average is 22.2. This observation points up the fact that if species-related errors are present, the choice of species to be included in the sample can have a pronounced effect on the estimation of sales-weighted shortfall. Similar biases could be introduced by inappropriate choice of data sources.

4.2.2 Some Formal Results

It is now appropriate to formalize some of the results of the previous section. These results pertain to the use of weighting factors if it is known that the regression of in-use fuel economy on EPA fuel economy is subject to species-related errors. First, however, there are some preliminaries.

In unweighted, least-squares regression analysis, one minimizes the sum of the squared residuals

$$\sum e_i^2 = \sum (y_i - b_0 - b_1 x_i)^2 = \text{minimum}$$

This requirement implies that the sum of the errors is zero, as can be shown by the following simple argument.

No loss of generality is involved if one references the x_i to their mean value $\bar{x}$ (see Interpretation Paper No. 2, Appendix II, op. cit.). The result is

$$b_0 = \sum y_i / n$$

$$b_1 = \frac{\sum (x_i - \bar{x}) y_i}{\sum (x_i - \bar{x})^2}$$

Then

$$\begin{aligned} \sum e_i &= \sum [y_i - b_0 - b_1 (x_i - \bar{x})] \\ &= \sum y_i - nb_0 - b_1 \sum (x_i - \bar{x}) = 0 \end{aligned}$$

since $\sum (x_i - \bar{x}) = 0$ and $nb_0 = \sum y_i$.

Similarly, in weighted least-squares analysis, one minimizes the weighted sum of squared residuals

$$\sum w_i e_i^2 = \sum w_i (y_i - b_0 - b_1 x_i)^2 = \text{minimum.}$$

Again, no loss of generality is involved if the x_i are referenced to their weighted mean value

$$\bar{x}_w = \frac{\sum w_i x_i}{\sum w_i},$$

in which case

$$b_0 = \frac{\sum w_i y_i}{\sum w_i}$$

and

$$b_1 = \frac{\sum w_i (x_i - \bar{x}_w) y_i}{\sum w_i (x_i - \bar{x}_w)^2}.$$

Then

$$\begin{aligned} \sum w_i e_i &= \sum w_i [y_i - b_0 - b_1 (x_i - \bar{x}_w)] \\ &= \sum w_i y_i - b_0 \sum w_i - b_1 \sum w_i (x_i - \bar{x}_w) = 0 \end{aligned}$$

since $\sum w_i (x_i - \bar{x}_w) = 0$ and since $b_0 \sum w_i = \sum w_i y_i$.

When regression analysis is used to compute the sales-weighted, average, in-use fuel economy, it is for the purpose of filling data gaps for species of vehicles represented in the population but not

represented in the sample. Accordingly, the averaging process is based on a weighted average of in-use fuel economy numbers taken from the regression line rather than actual, in-use fuel-economy numbers as measured. The implication is that equivalent averages will be obtained by the two procedures, even though measurements are not available on some species, or at least the regression-based average will not be subject to large error.

Figure 2 is provided to clarify this point. In the figure, there are three values of x , namely, 15, 20, 25. The corresponding observed values of y are 13, 20, 21. The unweighted mean of these values is $(13 + 20 + 21)/3 = 18$. Note that the fitted regression line yields three different values, 14, 18, 22, but having the same unweighted mean $(14 + 18 + 22)/3 = 18$. This result of course, is in accordance with the fact that the sum of the residuals around the regression line is zero. Suppose, however, one wishes to expand the average to include results for $x = 17.5$ and $x = 21.5$. Now no measured results are available, but it is presumed that one can substitute the corresponding y -values from the line, namely 16 for $x = 17.5$ and 19 for $x = 21.5$, with unweighted average

$$\bar{y} = (14 + 16 + 18 + 19.2 + 22)/5 = 17.84.$$

As noted in Interpretation Paper No. 2 (op. cit.), this is the same result as is obtained if the regression equation is evaluated at

$$\bar{x} = (15 + 17.5 + 20 + 21.5 + 25)/5 = 19.8.$$

The approach illustrated above is equally applicable in the event that a weighted least-squares analysis is performed. In fact, Figure 2 can be regarded as the result of a weighted analysis in which all three points are given equal weight. Suppose, however, that one now wishes to weight the three species as follows: 0.25, 0.50, 0.25. Applying these weights to the original observations, one obtains

$$(0.25)(13.0) + (0.50)(20.0) + (0.25)(21.0) = 19.5.$$

Applying these weights to the values computed from the regression equation, one obtains

$$(0.25)(14.0) + (0.50)(18.0) + (0.25)(22.0) = 18.0$$

FIGURE 2 A Hypothetical Fuel-Economy Regression Relation

Thus, unless the same weights are used in deriving the equation as are used in computing an average fuel economy, bias can result in estimation of the weighted mean.

It is at this point that a dilemma arises. Unless all species of vehicles are represented in the sample on which the regression is based, it is impossible to incorporate the appropriate weightings in the derivation of the regression equation. This fact is reflected in the earlier example in which two interpolated points are added to the three original points of Figure 2. The addition of these points yields weightings inconsistent with those assumed in the computation of original regression equation. For no bias to be introduced, it must be presumed that the weighted sum of the species-related errors for all five species, relative to the regression line, sum to zero.

5. Summary and Conclusions

Various error models for the regression analysis of in-use fuel economy have been considered. These admit primarily two error concepts: (1) random errors having zero expected value for all vehicle species, and (2) random errors having finite expected value peculiar to each vehicle species.

Under the first error assumption, the assumption of zero expected value for each species, the effects of error propagation can be addressed in accordance with the usual theory of regression analysis, as set forth in Interpretation Paper No. 2 (op. cit.). The following conclusions apply:

- The sales-weighted, average, in-use fuel-economy can be obtained by evaluating the regression equation for the sales-weighted average EPA number, with no loss in precision.
- So far as estimation of average shortfall is concerned, the selection of vehicles to be included in the data base is not critical, provided the average EPA rating for these vehicles is in the vicinity of the corresponding sales-weighted average for the population.
- For the evaluation of shortfall as a function of EPA-based fuel economy it is preferable to concentrate the vehicles in the data base at the extreme ends of the fuel-economy range of interest.

If a common error variance is applicable to all species, the sample weightings for each species should ideally be the same as the population weighting. More specifically, if the sales weight for a particular species is p percent of the population of all vehicles sold, then p percent of the regression sample should consist of that species. If the sample violates this rule, however, the regression equation can still provide an unbiased estimate of the population-weighted, in-use fuel economy. In computing the regression equation, it is appropriate to give each vehicle the same weight, regardless of the fact that the numbers of each species in the sample may not be in proportion to the number of vehicles of each species in the population.

Under the second error assumption, the assumption of a finite mean error peculiar to each species, the following conclusions apply:

- Inappropriate selection of vehicle species to be included in the regression sample can result in a biased estimate of the sales-weighted, average, in-use fuel economy.
- Ideally, all vehicle species present in the population should be represented in the sample; their proportions in the sample should be the same as their proportions in the population, and each vehicle should be given the same weight if homoscedasticity is assumed.
- If all species in the population are represented in the sample but not in the proportions specified above, then a weighted least-squares approach to the regression equation should be applied. Under the assumption of homoscedasticity, the weights for each vehicle species should be the same as their weights in the population, in order to eliminate possible bias in estimation of in-use, sales-weighted, average fuel economy.
- If all species in the population are not represented in the sample, no pronouncement can be made as to what weights should be used in computing the regression equation. It is desirable that the population-weighted sum of the species-related errors be zero, but this objective applies to all species, not just those which make up the sample, and there is no means by which this end can be assured, since the species-related errors are unknown.

Because of the impasse posed by the last conclusion, one must rely on engineering judgment and experience as a guide to assure that bias is eliminated to the greatest extent possible in any attempt to estimate sales-weighted, average, in-use fuel economy from an incomplete sample. Fortunately, there is reason to believe that species-related errors are small relative to vehicle-to-vehicle variation within species, unless some technology-specific factor, such as diesel engines, is known to apply. If such is the case, then it is recommended that separate regression equations be derived for the several technologies and that the results from these equations be combined by appropriate weighting based on known sales weights in the population. So long as the species-related errors are small, the assumption of zero expected value of the errors approximately apply, and the rules derived in Interpretation Paper No. 2 (op. cit.) provide the preferred approach to development of the regression equation.