BIOMETRICS 37, 391–411
June 1981

Building Multiple Regression Models Interactively

Harold V. Henderson
Ruakura Agricultural Research Centre, Hamilton, New Zealand

and

Paul F. Velleman
New York State School of Industrial and Labor Relations, Cornell University,
P.O. Box 1000, Ithaca, New York 14853, U.S.A.

SUMMARY

Automated multiple regression model-building techniques often hide important aspects of data from the data analyst. Such features as nonlinearity, collinearity, outliers, and points with high leverage can profoundly affect automated analyses, yet remain undetected. An alternative technique uses interactive computing and exploratory methods to discover unexpected features of the data. One important advantage of this approach is that the data analyst can use knowledge of the subject matter in the resolution of difficulties. The methods are illustrated with reanalyses of the two data sets used by Hocking (1976, *Biometrics* **32,** 1–44) to illustrate the use of automated regression methods.

1. Introduction

Practical multiple regression analysis has been changed radically in the past decade by the development and ready availability of automated regression techniques. Methods for building regression models are widely available in statistical packages and, as the cost of computing has dropped, they have become common data analysis tools. Yet many features of a data set can significantly alter the results of an automated regression and still remain undetected.

We propose an alternative philosophy of computer-assisted data analysis: a collaboration of the data analyst and the computer, made practicable by the growing availability of interactive computing. We illustrate this philosophy with specific techniques for building a multiple regression model, and with two examples. Our approach constrasts with such automated procedures as stepwise, forward selection, backwards elimination, best subsets and principal components regression. These and other related methods were described and illustrated by Hocking (1976). For direct comparison, we base our examples on the same two data sets employed by Hocking.

Our aim (like Hocking's) in analyzing these examples is not to discover the ultimate models for these data sets. We intend only to demonstrate that methods such as we propose are able to reveal features of data not found by automated methods, which change our understanding and lead to quite different analyses.

2. Exploring Regression Data

The fundamental axiom of this philosophy of data analysis is the declaration:

The data analyst knows more than the computer.

Key words: Data analysis; Outliers; Leverage; Stem-and-leaf display; Gasoline mileage; Air pollution and mortality.

More precisely, there is almost always information available to the data analyst (perhaps by asking the subject-discipline expert) that is not embodied in the numbers presented to the computer. In many cases the information could neither be effectively supplied to nor used by the computer. For example, most data analysts will know: (i) Water freezes at 0°C, so a discontinuity at that temperature might not be surprising. (ii) People tend to round their age to the nearest five-year boundary. (iii) The Arab oil embargo made 1973 an unusual year for many economic measures. (iv) The lab technician who handled Sample 3 is new on the job. (v) It is harder to weigh a brown bear in the forest than to measure its height and girth.

Many decisions made during a multivariate analysis (e.g. which variable to add next in a regression model) can involve selections among several alternatives which are almost equally good in the strict numerical sense of fitting the model to the data. In the absence of human guidance, the computer will make these decisions arbitrarily. In practice, a data analyst is likely to have a preference among the alternatives for other reasons, such as interpretability or relationship to theory. What may be worse is that the computer's choice will usually be optimal for some criterion (e.g. maximizing R^2 or minimizing C_p) and can thus be presented as objective.

Objectivity is a poor argument for automated analyses. The 'objective' decisions of the computer are tolerated only when the data analyst first makes the *subjective* decision to abdicate further responsibility.

Many features of the data can affect the analysis without arousing the suspicions of the researcher or data analyst. Consequently, it is often better to examine the appropriate display than to compute the appropriate statistics. Displays accommodate the unexpected more gracefully. By 'the unexpected' we mean both the unexpected features we are taught to expect (nonlinearity, outliers, heteroskedasticity) and less easily described features such as discontinuities in a trend, or the clumping of observations.

3. Methods

Many of the techniques we illustrate here for building a regression model interactively have been presented elsewhere, although not always in this context. One excellent source is the book by Mosteller and Tukey (1977), which we recommend highly to anyone who computes or reads regression analyses.

3.1 *Univariate Data Display*

To examine univariate distributions we use stem-and-leaf displays (Tukey, 1977). They have most of the virtures of histograms and two great advantages: they identify individual points, and they are not rendered useless by extraordinary values. Velleman and Hoaglin (1981) provided computer programs for stem-and-leaf displays and specified a display-scaling algorithm that resists the effects of extraordinary points.

3.2 *Transforming Data*

Data re-expressions can often make an asymmetric distribution more symmetric or a curved x–y relationship more nearly linear. These two applications of data re-expression are considered together by one technique for finding a good re-expression.

For univariate data, define y_A = the qth fractile of the data, $y_B = y_C$ = the median, and y_D = the $(1-q)$th fractile. If the data are symmetric, $(y_B - y_A)/(y_D - y_C) = 1$ at each fractile. For an x–y relationship, define (x_A, y_A), $(x_B, y_B) = (x_C, y_C)$ and (x_D, y_D) to be three

points (not necessarily in the data) which summarize each of three partitions of the x–y plane determined by dividing the x-axis into three regions, roughly equally spaced. [Velleman and Hoaglin (1981) gave an algorithm and computer program to find suitable points.] For simplicity, we assume equal spacing: $x_B - x_A = x_D - x_C = \Delta x$. If the x–y relationship is linear, the two 'half slopes' $b_1 = (y_B - y_A)/\Delta x$ and $b_2 = (y_D - y_C)/\Delta x$ will be equal. Thus, in both cases the closeness of the data structure to the desired structure (symmetry or linearity) can be measured approximately by the closeness of

$$h = \log\{(y_B - y_A)/(y_D - y_C)\} \tag{3.1}$$

to zero.

The effect of the re-expression $z = f(y)$ can be measured by the change in h:

$$\begin{aligned} h(f) - h(1) &= \log\{(z_B - z_A)/(z_D - z_C)\} - \log\{(y_B - y_A)/(y_D - y_C)\} \\ &= \log\{(z_B - z_A)/(y_B - y_A)\} - \log\{(z_D - z_C)/(y_D - y_C)\}. \end{aligned} \tag{3.2}$$

Now, allowing the summary points and their transformations to converge according to $y_A \to y_1 \leftarrow y_B$ and $y_C \to y_2 \leftarrow y_D$,

$$h(f) - h(1) = \log\{dz/dy\}|_{y_1} - \log\{dz/dy\}|_{y_2}. \tag{3.3}$$

For the common family of re-expressions,

$$z = f(y) = \begin{cases} y^p & p \neq 0, \\ \log(y) & p = 0, \end{cases} \tag{3.4}$$

(3.3) reduces to

$$(p-1)\log(y_1/y_2). \tag{3.5}$$

The closeness of the data to the desired structure, as measured by $h(p)$, is thus roughly linear in the power chosen, with a slope of

$$\frac{h(1) - h(p)}{1 - p}. \tag{3.6}$$

The use of this slope for one step of Newton's method yields a proposed power:

$$\hat{p} = \frac{ph(1) - h(p)}{h(1) - h(p)}, \tag{3.7}$$

which should do a better job of simplifying the data structure. We could iterate (3.7) but we usually prefer an integer or simple rational power, so we are likely to round the first or second estimate to a nearby simple value rather than iterate to convergence.

3.3 *Extraordinary Data Points*

Many writers have discussed the detection and treatment of outliers in y. We use stem-and-leaf displays of residuals, quantile–quantile plots of the ordered values of y against expected Gaussian values (Wilk and Gnanadesikan, 1968), and partial regression plots (see below) to locate outliers and informally assess their influence, but many other techniques can be useful.

Points which are extraordinary in predictor space can also have a large and unexpected influence on regression analyses and on automated regression procedures. Such points can be detected by their large 'leverage': the leverage of the ith data point in the multiple

regression, $\mathbf{y}=\mathbf{X}\boldsymbol{\beta}+\mathbf{e}$, is the ith diagonal element, h_i, of the projection matrix, $\mathbf{H}=\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. Hoaglin and Welsch (1978) discussed the properties of $\mathbf{H}$; they showed that $0\leq h_i\leq 1$ (where $h_i=1$ corresponds to devoting a degree of freedom to fitting the model to the ith point), and that replacement of y_i by y_i+1 in the regression changes the corresponding predicted value $\hat{y}_i$ to $\hat{y}_i+h_i$. They suggested that in a regression with p carriers and n data points, leverage values greater than $2p/n$ deserve attention as possibly extreme. In particular, the leverage of the ith point in a regression with one predictor is $(x_i-\bar{x})^2/\sum_k (x_k-\bar{x})^2$. The leverage of the ith point h_i, and its residual, e_i, are related by

$$h_i+e_i^2\Big/\sum_k e_k^2\leq 1. \tag{3.8}$$

Cook (1977) measured the influence of each data value on the vector of coefficients in terms of the Mahalanobis distance that $\hat{\boldsymbol{\beta}}$ would move were the ith data point omitted. His measure can be written as

$$D_i=e_i^2/\{ps^2(1-h_i)^2\}, \tag{3.9}$$

where p is the number of coefficients estimated and s^2 is the usual estimate of the variance of the regression residuals. This statistic can fail to identify influential points if several extraordinary points support each other. We prefer to examine leverage and residuals separately.

3.4 *Partial Regression Plots*

Partial regression plots, used by Mosteller and Tukey (1977) and discussed by Belsley, Kuh and Welsch (1980), are a fundamental tool of interactive regression model building. For the multiple regression of y on 'carriers'[1] $x_0\equiv 1, x_1, \ldots, x_p$, i.e.

$$y=b_0+b_1x_1+\cdots+b_px_p+e, \tag{3.10}$$

define $y_{\cdot 23\ldots p}$ (or, more precisely, $y_{\cdot 023\ldots p}$) to be the residuals from the least squares regression of y on all carriers except x_1, and define $x_{1\cdot 23\ldots p}$ similarly as the residuals from the regression of x_1 on the other carriers. The partial regression plot for b_1 is the plot of $y_{\cdot 23\ldots p}$ against $x_{1\cdot 23\ldots p}$. It has least squares slope b_1 and least squares residuals equal to the final residuals, e, and it displays the effects of individual points on the estimation of b_1 in the full regression. The correlation of $y_{\cdot 23\ldots p}$ and $x_{1\cdot 23\ldots p}$ is a partial correlation. We write $x_{j\cdot\text{rest}}$ to indicate regression on the rest of the carriers excluding x_j. Note that the final residuals, $e=y_{\cdot 012\ldots p}$.

The partial regression plot also depicts the partial leverage, $\eta_{j\cdot\text{rest}}(i)$, of the ith data point due to the jth carrier. Specifically, $\eta_{j\cdot\text{rest}}(i)$ is the amount by which h_i would increase (decrease) were x_j added to (removed from) the regression model. Belsley *et al.* (1980) showed that

$$\eta_{j\cdot\text{rest}}(i)=\frac{x_{j\cdot\text{rest}}^2(i)}{\sum_k x_{j\cdot\text{rest}}^2(k)}. \tag{3.11}$$

Because $x_{j\cdot\text{rest}}$ has zero mean, (3.11) is the leverage of the ith point in a one-carrier regression on $x_{j\cdot\text{rest}}$—that is, the leverage of the ith point in the partial regression plot for b_j.

---

[1] We follow the terminology introduced by Mosteller and Tukey (1977) and refer to the x_j (including the constant x_0) as 'carriers'. A similar usage was followed by Sylvester (1884) in connection with simultaneous equations.

Partial regression plots are similar to a diagnostic display discussed by Ezekiel (1924), Larsen and McCleary (1972), and Wood (1973). Their display plots

$$e + b_j x_j \quad \text{against} \quad x_j, \tag{3.12}$$

where b_j is estimated from the full multiple regression. Mosteller and Tukey (1977) showed that partial regression plots are equivalent to plotting

$$e + b_j x_{j\cdot\text{rest}} \quad \text{against} \quad x_{j\cdot\text{rest}}. \tag{3.13}$$

We find both displays useful. The form of (3.12) is especially valuable in discovering a useful re-expression of x_j which will linearize its relationship with y. Nevertheless, we generally prefer the partial regression plot (3.13), and use (3.12) only when a partial regression plot indicates the need to re-express x_j. Partial regression plots help the data analyst to recognize the inter-relationships among the predictors in the model. For example, the differences between, say, $x_{1\cdot 23}$ and $x_{1\cdot 234}$ may be of greater magnitude and importance than the corresponding differences between $y_{\cdot 23}$ and $y_{\cdot 234}$, and so may better account for the changes in b_1 brought about by introduction of x_4.

3.5 *Inference and Validation*

Clearly, interactive regression model building is far enough removed from classical regression theory that we should be uneasy about using traditional procedures of statistical inference. Nevertheless, such measures as t statistics for coefficients, multiple R^2 (adjusted for degress of freedom), and the F statistic for the regression are still useful measures of the success of a model in fitting a data set. Often a good fit is all that is needed. [Mosteller and Tukey (1977) discussed the value of asking no more from an analysis than an indication of the patterns in the data.]

Where enough data is available, it may be wise to split the data in half randomly, build a model using one half and examine the success of the model on the other half. Such validation can, of course, be done in both directions and with more or different size partitions. At other times additional data may be available from a subsequent study.

3.6 *Combining the Techniques*

In the examples below we employ these techniques to build regression models. While we neither recommend nor follow a specific paradigm, we outline here some ways in which we have coordinated the various techniques. We begin by using stem-and-leaf displays and related exploratory methods (Tukey, 1977; Velleman and Hoaglin, 1981) to identify skewed univariate distributions, multimodalities, possible bad data, and other extraordinary patterns or items. We plot y against each available carrier to discover simple nonlinear relationships. On the basis of these displays, we re-express variables in order to improve symmetry and linearity where this seems advisable. While it is still possible for some variables to benefit from re-expression which will simplify multivariate relationships, it is well worth the effort to do the simple univariate and bivariate checks first.

In building the regression models we use correlation and partial correlation (of $y_{\cdot\text{carriers in model}}$ with candidate carriers) to nominate carriers for consideration. We also compute maximum partial leverage and maximum absolute residuals to identify those cases where the partial correlations might have been altered by extraordinary points. Partial regression plots are produced for each nominated carrier and examined in order to identify influential data points and to select variables to add to the model. We examine

the final regression model by computing and analyzing residuals, computing tolerance values for each carrier, and examining any omitted points.

Most computations were performed interactively on the Minitab statistics package (Ryan, Joiner and Ryan, 1976, 1980) on Cornell University's IBM 370/168 under a modified CMS operating system. Automated regressions were computed on SAS (Helwig and Council, 1979) and BMDP (Dixon and Brown, 1979). Some exhibits were generated with the programs by Vellemen and Hoaglin (1981) incorporated into Minitab.

4. Gasoline Mileage Data

The data, extracted from 1974 *Motor Trend* magazine, comprise gasoline mileage in miles per gallon (MPG), and ten aspects of automobile design and performance (Table 1) for 32 automobiles (1973–74 models). The regression model attempts to predict gasoline mileage

Table 1
*Data for Motor Trend sample of 32 automobiles**

Code†	Automobile	MPG	CYL	DISP	HP	DRAT	WT	QSEC	V/S	A/M	GEAR	CARB
A	Mazda RX-4‡	21.0	6	160.0	110	3.90	2.620	16.46	0	1	4	4
B	Mazda RX-4 Wagon‡	21.0	6	160.0	110	3.90	2.875	17.02	0	1	4	4
C	Datsun 710	22.8	4	108.0	93	3.85	2.320	18.61	1	1	4	1
D	Hornet 4 Drive	21.4	6	258.0	110	3.08	3.215	19.44	1	0	3	1
E	Hornet Sportabout	18.7	8	360.0	175	3.15	3.440	17.02	0	0	3	2
F	Valiant	18.1	6	225.0	105	2.76	3.460	20.22	1	0	3	1
G	Duster 360	14.3	8	360.0	245	3.21	3.570	15.84	0	0	3	4
H	Mercedes 240D‡	24.4	4	146.7	62	3.69	3.190	20.00	1	0	4	2
I	Mercedes 230	22.8	4	140.8	95	3.92	3.150	22.90	1	0	4	2
J	Mercedes 280	19.2	6	167.6	123	3.92	3.440	18.30	1	0	4	4
K	Mercedes 280C	17.8	6	167.6	123	3.92	3.440	18.90	1	0	4	4
L	Mercedes 450SE	16.4	8	275.8	180	3.07	4.070	17.40	0	0	3	3
M	Mercedes 450SL	17.3	8	275.8	180	3.07	3.730	17.60	0	0	3	3
N	Mercedes 450SLC	15.2	8	275.8	180	3.07	3.780	18.00	0	0	3	3
O	Cadillac Fleetwood	10.4	8	472.0	205	2.93	5.250	17.98	0	0	3	4
P	Lincoln Continental	10.4	8	460.0	215	3.00	5.424	17.82	0	0	3	4
Q	Chrysler Imperial	14.7	8	440.0	230	3.23	5.345	17.42	0	0	3	4
R	Fiat 128	32.4	4	78.7	66	4.08	2.200	19.47	1	1	4	1
S	Honda Civic	30.4	4	75.7	52	4.93	1.615	18.52	1	1	4	2
T	Toyota Corolla	33.9	4	71.1	65	4.22	1.835	19.90	1	1	4	1
U	Toyota Corona	21.5	4	120.1	97	3.70	2.465	20.01	1	0	3	1
V	Dodge Challenger	15.5	8	318.0	150	2.76	3.520	16.87	0	0	3	2
W	AMC Javelin	15.2	8	304.0	150	3.15	3.435	17.30	0	0	3	2
X	Camaro Z-28	13.3	8	350.0	245	3.73	3.840	15.41	0	0	3	4
Y	Pontiac Firebird	19.2	8	400.0	175	3.08	3.845	17.05	0	0	3	2
Z	Fiat X1-9	27.3	4	79.0	66	4.08	1.935	18.90	1	1	4	1
A	Porsche 914-2‡	26.0	4	120.3	91	4.43	2.140	16.70	0	1	5	2
B	Lotus Europa	30.4	4	95.1	113	3.77	1.513	16.90	1	1	5	2
C	Ford Pantera L	15.8	8	351.0	264	4.22	3.170	14.50	0	1	5	4
D	Ferrari Dino 1973	19.7	6	145.0	175	3.62	2.770	15.50	0	1	5	6
E	Maserati Bora	15.0	8	301.0	335	3.54	3.570	14.60	0	1	5	8
F	Volvo 142E	21.4	4	121.0	109	4.11	2.780	18.60	1	1	4	2

* By courtesy of Dr R. R. Hocking.

† These letters are used for identification in Figs.

‡ Hocking's noncrucial coding of the Mazda's rotary engine as a straight six-cylinder engine and the Porsche's flat engine as a V engine, as well as the inclusion of the diesel Mercedes 240D, have been retained to enable direct comparisons to be made with previous analyses.

from the other variables. As Table 1 shows, this particular sample of 32 automobiles has a bias to exotic, non-U.S., automobiles: it includes seven Mercedes, a Porsche, a Ferrari and a Maserati. Therefore, we might not expect a universal prediction model to emerge.

Hocking (1976) analysed these data using various stepwise regression techniques, all-subset regressions with a variety of subset selection criteria such as minimum C_p and minimum residual mean square (RMS), the biased estimation techniques of Stein shrinkage and ridge regression, and principal component regression. On the basis of these analyses he suggested (p. 12) that the regression of MPG on TRANSMISSION TYPE, WEIGHT, and QUARTER MILE TIME 'may be best for prediction'. This model is difficult to interpret, and the absence of DISPLACEMENT or HORSEPOWER, which intuition suggests should be important in the prediction of MPG, is surprising. Hocking noted (pp. 25–26) that DISPLACEMENT exhibits instability and he found it disturbing that the ridge trace did not suggest the important role of his optimal subset.

Multicollinearity aside, disturbing features like these are not uncommon in blind automated analyses. Simple inspection of plots of MPG on the carriers reveals curved relationships with several important variables, including WEIGHT and DISPLACEMENT (Fig. 1). The iteration in (3.7) suggests that $(\text{MPG})^{-1}$ (= gallons per mile, GPM) would be more nearly linear against these carriers. (The metric equivalent, liters per 100 kilometers, is in fact used in many countries.) To obtain more convenient units we rescaled GPM to gallons per 100 miles. Figure 2 shows improvement. GPM is also more likely to satisfy additive and homogeneous error assumptions because fuel consumption was measured over a fixed 73-mile route.

Correlations and plots of GPM against promising carriers suggest WEIGHT as the best single predictor of GPM. There is also a loose theoretical argument that adds credence to this model. If gasoline consumed is proportional to the work expended in moving the vehicle [= force × distance], then on a per-mile basis, GPM ∝ force ∝ WEIGHT. The regression of GPM on WEIGHT yielded the equation GPM = 0.617 + 1.49 WEIGHT (GPM in gallons per 100 miles, WEIGHT in 1000 lbs), and marked the three U.S. luxury cars as high-leverage points in this sample.

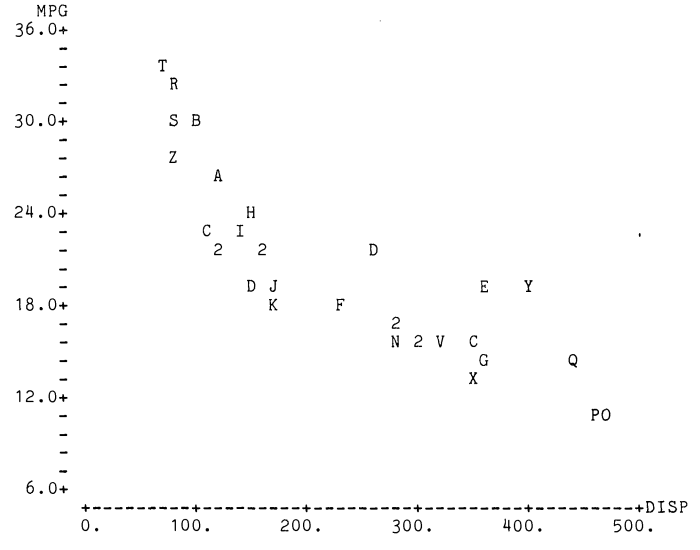

Figure 1. MPG vs DISPlacement. Note the nonlinear relationship. (See Table 1 to identify automobiles with letters.)

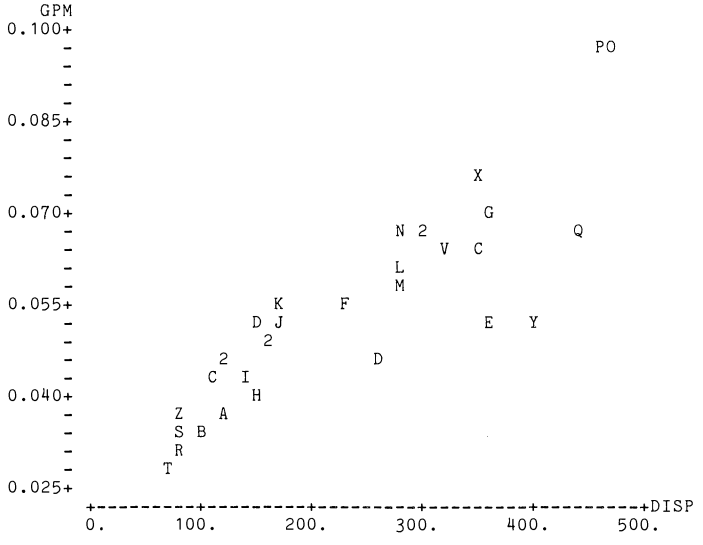

Figure 2. GPM vs DISPlacement. Note the straightening of the relationship due to the re-expression of MPG.

At this point we might want to include DISPLACEMENT or HORSEPOWER, but these variables offer little improvement in the model (possibly due to strong correlations with WEIGHT). An understanding of the types of cars in this collection suggests that a measure of how overpowered a car is might be useful. We computed the ratio of horsepower to weight, HP/WT, which is almost uncorrelated with WEIGHT ($r = .003$), and offered it as a carrier. It is the most promising contender for inclusion in the model at this step, having a partial correlation of 0.52 with GPM · WT. The partial regression plot of GPM · WT on HP/WT · WT (Fig. 3) shows two roughly parallel bands of points with

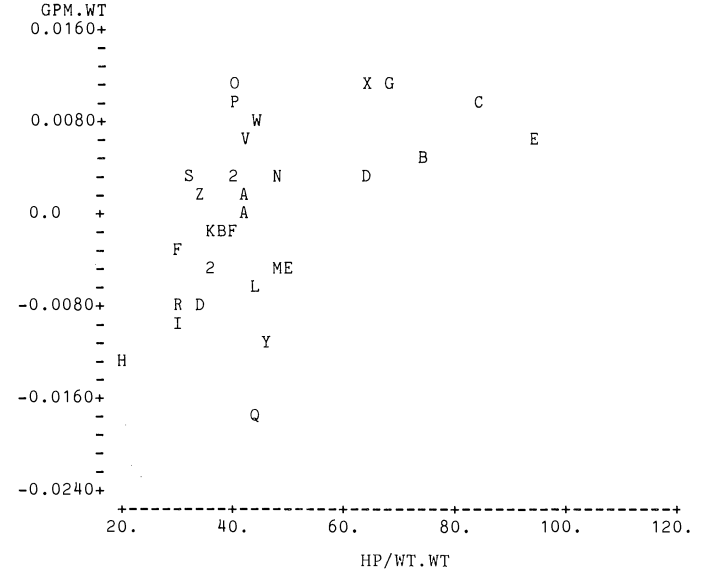

Figure 3. Partial regression plot of GPM · WT vs HP/WT · WT. Note the sports cars clustered in the upper right.

Table 2
Regression of GPM on WT and HP/WT for Motor Trend sample

	Column	Coefficient	SD of coefficient	t ratio = coeff./SD
		−0.401	0.512	−0.78
x_1	WT	1.472	0.122	12.11
x_2	HP/WT	0.02400	0.00730	3.29

SD of y about regression: $s = 0.0661$ with 32-3 = 29 df.
$R^2 = 84.8\%$
$R^2 = 83.8\%$, adjusted for df.

	Row	y GPM	Predicted y	SD of pred. y	Residual	Stand. resid.
Cadillac Fleetwood	15	9.62	8.26	0.28	1.35	2.25R
Lincoln Continental	16	9.62	8.53	0.30	1.08	1.83X
Chrysler Imperial	17	6.80	8.50	0.29	−1.70	−2.84R
Lotus Europa	28	3.29	3.62	0.33	−0.33	−0.57X
Ford Pantera L	29	6.33	6.26	0.30	0.07	0.11X
Maserati Bora	31	6.67	7.11	0.37	−0.44	−0.08X

R denotes an observation with a large standardized residual.
X denotes an observation whose x value gives it large influence.

most high-performance cars in the band to the right. This is the kind of unexpected pattern best revealed by displays. The gap between the bands might be due more to the omission of midsize American cars from this sample of automobiles than to an anomaly of automobile performance.

Table 2 details the regression of GPM on WEIGHT and HP/WT and also notes those points with leverages greater than .19 ($=2p/n$) or with large residuals. The remaining available carriers do not significantly improve this model. In fact, we do slightly better by

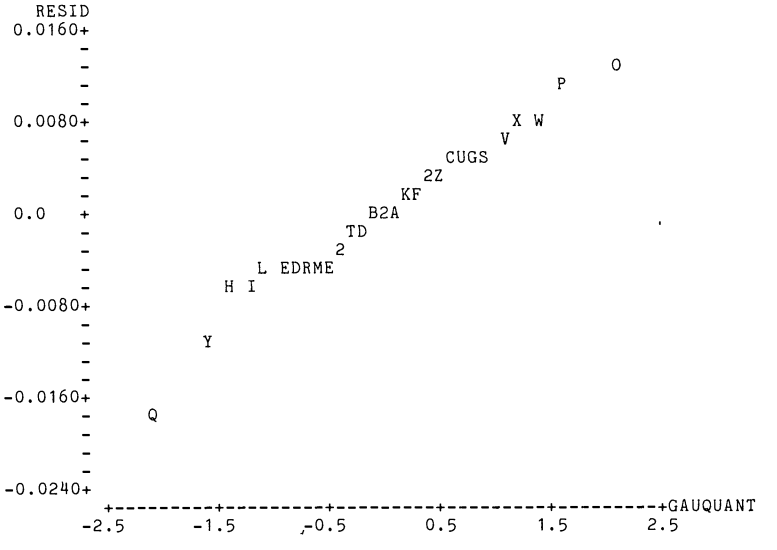

Figure 4. Quantile–quantile (Q–Q) plot of residuals from the regression of Table 2.

fitting through the origin, which may make sense theoretically. The individual t statistics of the estimated coefficients are solid (12.11 and 3.29) and the residual Gaussian Q–Q plot (Fig. 4) is reasonably straight. In all, this three-carrier model predicts as well as Hocking's four-carrier automated regression solution, but is much easier to understand and interpret.

To validate the model we collected data on 38 automobiles (1978–79 models) from *Consumer Reports*. Their MPG measure is similar to that in *Motor Trend*, being based in this case on a 123-mile test drive. Most other variables (including HORSEPOWER and WEIGHT) are recorded as reported by the manufacturers. This sample (Table 3) is more homogeneous than the *Motor Trend* sample, including American family sedans and station wagons as well as compacts and subcompacts.

Table 3
Data for Consumer Reports sample of 38 automobiles

Code	Automobile	MPG	CYL	DISP	HP	DRAT	WT	ACCEL	ENGTYPE
A	Buick Estate Wagon	16.9	8	350	155	2.73	4.360	14.9	1
B	Ford Country Sq. Wagon	15.5	8	351	142	2.26	4.054	14.3	1
C	Chevy Malibu Wagon	19.2	8	267	125	2.56	3.605	15.0	1
D	Chrys. Lebaron Wagon	18.5	8	360	150	2.45	3.940	13.0	1
E	Chevette	30.0	4	98	68	3.70	2.155	16.5	0
F	Toyota Corona	27.5	4	134	95	3.05	2.560	14.2	0
G	Datsun 510	27.2	4	119	97	3.54	2.300	14.7	0
H	Dodge Omni	30.9	4	105	75	3.37	2.230	14.5	0
I	Audi 5000	20.3	5	131	103	3.90	2.830	15.9	0
J	Volvo 240 GL	17.0	6	163	125	3.50	3.140	13.6	0
K	Saab 99 GLE	21.6	4	121	115	3.77	2.795	15.7	0
L	Peugeot 694 SL	16.2	6	163	133	3.58	3.410	15.8	0
M	Buick Century Spec.	20.6	6	231	105	2.73	3.380	15.8	0
N	Mercury Zephyr	20.8	6	200	85	3.08	3.070	16.7	0
O	Dodge Aspen	18.6	6	225	110	2.71	3.620	18.7	0
P	AMC Concord D/L	18.1	6	258	120	2.73	3.410	15.1	0
Q	Chevy Caprice Classic	17.0	8	305	130	2.41	3.840	15.4	1
R	Ford LTD	17.6	8	302	129	2.26	3.725	13.4	1
S	Mercury Grand Marquis	16.5	8	351	138	2.26	3.955	13.2	1
T	Dodge St Regis	18.2	8	318	135	2.45	3.830	15.2	1
U	Ford Mustang 4	26.5	4	140	88	3.08	2.585	14.4	0
V	Ford Mustang Ghia	21.9	6	171	109	3.08	2.910	16.6	1
W	Mazda GLC	34.1	4	86	65	3.73	1.975	15.2	0
X	Dodge Colt	35.1	4	98	80	2.97	1.915	14.4	0
Y	AMC Spirit	27.4	4	121	80	3.08	2.670	15.0	0
Z	VW Scirocco	31.5	4	89	71	3.78	1.990	14.9	0
A	Honda Accord LX	29.5	4	98	68	3.05	2.135	16.6	0
B	Buick Skylark	28.4	4	151	90	2.53	2.670	16.0	0
C	Chevy Citation	28.8	6	173	115	2.69	2.595	11.3	1
D	Olds Omega	26.8	6	173	115	2.84	2.700	12.9	1
E	Pontiac Phoenix	33.5	4	151	90	2.69	2.556	13.2	0
F	Plymouth Horizon	34.2	4	105	70	3.37	2.200	13.2	0
G	Datsun 210	31.8	4	85	65	3.70	2.020	19.2	0
H	Fiat Strada	37.3	4	91	69	3.10	2.130	14.7	0
I	VW Dasher	30.5	4	97	78	3.70	2.190	14.1	0
J	Datsun 810	22.0	6	146	97	3.70	2.815	14.5	0
K	BMW 320i	21.5	4	121	110	3.64	2.600	12.8	0
L	VW Rabbit	31.9	4	89	71	3.78	1.925	14.0	0

unit = 0.1
1 2 = 1.2

Original Results		Prediction Errors
6	−1·	
1	−1*	02
5556	−0·	5555668899
12334444	−0*	001111112222333444
444322100000	+0·	00022
8765	+0·	67
30	+1*	

Figure 5. Back-to-back stem-and-leaf display comparing residuals from original regression with prediction errors from the *Consumer Reports* validation data set.

We first predicted GPM for these cars using the model of Table 2, and then computed prediction errors (Fig. 5). The model did quite well: the mean square prediction error (0.313) is smaller than the original mean square residual in the *Motor Trend* data set (0.437). However, this may be due in part to the greater homogeneity of the *Consumer Reports* sample.

We then computed the regression of GPM on WEIGHT and HP/WT for the *Consumer Reports* data (Table 4). The coefficients of both carriers agree quite closely with those for the original model. However, the standard deviation of the HP/WT coefficient is more than twice as large as before.

Table 4
Regression of GPM on WT and HP/WT for Consumer Reports sample

	Column	Coefficient	SD of coefficient	t ratio = coeff./SD
	—	−1.0080	0.7331	−1.37
x_1	WT	1.523	0.101	15.06
x_2	HP/WT	0.0275	0.0184	1.50

SD of y about regression: $s = 0.434$ with 38-3 = 35 df.
$R^2 = 86.7\%$
$R^2 = 85.9\%$, adjusted for df.

	Row	y GPM	Predicted y	SD of pred. y	Residual	Stand. resid.
Volvo 240GL	10	5.88	4.86	0.10	1.01	2.41R
Peugeot 604SL	12	6.17	5.25	0.11	0.91	2.18R
Chevrolet Citation	29	3.47	4.16	0.17	−0.69	−1.74X
Pontiac Phoenix	31	2.98	3.85	0.07	−0.86	−2.03R

R denotes an observation with a large standardized residual.
X denotes an observation whose x value gives it large influence.

We conclude from this that the re-expressed variable GPM is indeed approximately linear in WEIGHT and HP/WT and that the equation GPM≐1.5 WEIGHT+.026 HP/WT seems useful for a wide variety of automobiles. The carrier HP/WT may be needed primarily to account for overpowered sports cars absent from the *Consumer Reports* sample. The intercept term may be useful, but it is difficult to tell from the samples. Certainly the linear extrapolation from 2000 lbs to the origin is dangerous. Perhaps data on motorcycles and mopeds might confirm or reject the linearity of the relationship near the origin.

5. Air Pollution and Mortality

McDonald and Schwing (1973) built a multiple regression model to predict age-adjusted mortality in 60 U.S. Standard Metropolitan Statistical Areas (SMSAs), from 15 variables measuring socio-economic, climatological, and air pollution features. Table 5 names the variables. They fitted regression models for the entire set of predictors and selected smaller models using an all-subsets selection method which minimized C_p (Mallows, 1973), and a 'ridge elimination' technique proposed by Hoerl and Kennard (1970). McDonald and Ayers (1978) employed multivariate graphical methods for these data (e.g. 'faces' proposed by Chernoff, 1973) and presented the results. Hocking (1976) reanalysed the data using all of the methods mentioned in §4. Table 6 shows some of the models proposed by these authors.

We first examined stem-and-leaf displays of all 16 variables. The most striking phenomenon is the extraordinary magnitude of air pollution in the California SMSAs. Los Angeles has a hydrocarbon potential of 648, while Chicago, the most extreme non-California city, has a value of 88. The extreme California air pollution values are especially hazardous to blind automated regression techniques because they are not isolated values. Los Angeles has the remarkable leverage of .907 in the 16-carrier model. Nevertheless, Cook's distance fails to flag any of these points as extraordinary, partly because their residuals are not especially large, and partly because the four California SMSAs support each other. That is, omitting any *one* of them would make little difference in the regression coefficients. Their effect on automated variable-selection computations is similarly disguised by the fact that they are extreme on all three air pollution variables.

The univariate distributions of the air pollution potentials are made nearly symmetric by re-expressing to logarithms, (e.g. Fig. 6), and plots of mortality against each of them become more nearly straight. [McDonald and Schwing (1973) suggested, but did not use, logarithmic re-expressions of the air pollution potentials; McDonald and Ayers (1979) did use this re-expression.] We added LOGHC, LOGNOX, and LOGSO2 to the data set. When the re-expressed variables replaced the raw air pollution potentials the (16-carrier) leverage of Los Angeles dropped to .53, behind the leverage of both Miami (.69) and New Orleans (.60).

The stem-and-leaf display of population density also aroused our suspicions. Two values are extreme: New York at 7462 people/mi^2 and York, Pennsylvania at 9699 people/mi^2. We checked to confirm that the latter value is correct; it appears to have been caused by an odd definition of the SMSA boundary for York. Even so, it hardly seems to reflect accurately the urban character of York. We resolved to identify York in all future diagnostic displays and to consider omitting it if it exerted an undue influence on the model.

We next examined the full correlation matrix. It suggested %NONWHITE as a good predictor ($r^2 = .414$), and a plot confirms this. MORTALITY was regressed on

Table 5
Description of variables, and list of cities in data set from McDonald and Schwing (1973)

Variable name	Description
RAIN	Mean annual precipitation in inches.
JAN	Mean January temperature in degrees Fahrenheit.
JULY	Mean July temperature in degrees Fahrenheit.
OLD	Percentage of 1960 SMSA population which is 65 years of age or over.
POP/HSE	Population per household, 1960 SMSA.
EDUC	Median school years completed for those over 25 in 1960 SMSA.
GDHSE	Percentage of housing units that are sound with all facilities.
POPDEN	Population per square mile in urbanized area of SMSA in 1960.
NONW	Percentage of 1960 urbanized area population that is nonwhite.
WCOL	Percent employment in white-collar occupations in 1960 urbanized area.
POOR	Percentage of families with incomes under $3000 in 1960 urbanized area.
HC	Relative pollution potential of hydrocarbons, HC.
NOX	Relative pollution potential of oxides of nitrogen, NOx.
SO2	Relative pollution potential of sulfur dioxide, SO_2.
HUMID	Percent relative humidity, annual average at 1 p.m.
MORT	Total age-adjusted mortality rate, expressed as deaths per 100 000 population.

Code	City	Code	City
A	Akron, Oh.	E	Memphis, Tn.
B	Albany, N.Y.	F	Miami, Fl.
C	Allentown, Pa	G	Milwaukee, Wi.
D	Atlanta, Ga	H	Minneapolis, Mn.
E	Baltimore, Md	I	Nashville, Tn.
F	Birmingham, Al.	J	New Haven, Ct
G	Boston, Ma.	K	New Orleans, La
H	Bridgeport, Ct	L	New York, N.Y.
I	Buffalo, N.Y.	M	Philadelphia, Pa
J	Canton, Oh.	N	Pittsburgh, Pa
K	Chattanooga, Tn.	O	Portland, Or.
L	Chicago, Il.	P	Providence, R.I.
M	Cincinnati, Oh.	Q	Reading, Pa
N	Cleveland, Oh.	R	Richmond, Va
O	Columbus, Oh.	S	Rochester, N.Y.
P	Dallas, Tx.	T	St. Louis, Mo.
Q	Dayton, Oh.	U	San Diego, Ca.
R	Denver, Co.	V	San Francisco, Ca.
S	Detroit, Mi.	W	San Jose, Ca.
T	Flint, Mi.	X	Seattle, Wa.
U	Fort Worth, Tx.	Y	Springfield, Ma.
V	Grand Rapids, Mi.	Z	Syracuse, N.Y.
W	Greensboro, N.C.	A	Toledo, Oh.
X	Hartford, Ct	B	Utica, N.Y.
Y	Houston, Tx.	C	Washington, D.C.
Z	Indianapolis, In.	D	Wichita, Ks
A	Kansas City, Mo.	E	Wilmington, De.
B	Lancaster, Pa	F	Worcester, Ma.
C	Los Angeles, Ca.	G	York, Pa
D	Louisville, Ky	H	Youngstown, Oh.

Table 6

Coefficients (t statistics) for various regression models for air pollution data

	Full models		Min C_p	Ridge elim.	Best subset $p = 8$	Our choice
CONST	1762 (4.03)	2069 (4.78)	120 (9.80)	788.4 (11.16)	1929 (6.62)	926.3 (11.70)
RAIN	1.905 (2.06)	2.905 (3.41)	1.797 (3.0)	1.487 (2.53)	1.648 (2.73)	2.0088 (4.45)
JAN	−1.95 (−1.76)	−3.20 (−3.08)	−1.484 (−2.84)	−1.633 (−3.15)	−1.893 (−3.21)	—
JULY	−3.10 (−1.63)	−3.87 (−1.87)	−2.36 (−1.89)	—	−2.30 (−1.86)	—
OLD	−9.07 (−1.07)	−15.01 (−1.93)	—	—	—	—
POP/HSE	−106.9 (−1.53)	−159.4 (−2.40)	—	—	−62.0 (−1.39)	—
EDUC	−17.2 (−1.45)	−18.9 (−1.82)	−13.62 (−2.12)	−11.53 (−1.74)	−16.97 (−2.40)	−16.89 (−2.97)
GDHSE	−0.62 (−0.35)	−0.58 (−0.37)	—	—	—	—
POPDEN	0.0036 (0.89)	0.00358 (0.93)	—	0.00415 (1.14)	—	.01046 (3.02)
NONW	4.46 (3.36)	4.00 (3.13)	4.585 (6.59)	4.144 (6.32)	5.216 (6.31)	2.467 (4.89)
WCOL	−0.18 (−0.11)	0.01 (0.01)	—	—	—	—
POOR	−0.14 (−0.04)	0.87 (0.30)	—	—	—	—
HUMID	0.11 (0.09)	−0.18 (−0.17)	—	—	—	—
HC	−0.673 (−1.37)	—	—	—	—	—
NOx	1.34 (1.34)	—	—	—	—	—
SO2	0.086 (0.58)	—	0.260 (3.31)	0.245 (2.86)	0.225 (2.76)	—
LOGHC	—	−81.6 (−2.32)	—	—	—	—
LOGNOX	—	124.3 (3.56)	—	—	—	29.12 (3.38)
LOGSO2	—	−18.7 (−1.16)	—	—	—	
YORK	—		—	—	—	−128.8 (−3.69)
LANCASTER	—	—	—	—	—	−113.9 (−4.00)
MIAMI	—	—	—	—	—	−104.4 (3.35)
NEW ORLEANS	—	—	—	—	—	82.1 (2.86)
Adj. R^2	68.5%	72.8%	70.5%	69.3%	71.0%	81.6%
F	9.55	11.55	24.48	25.20	21.6	30.17
s(df)	34.9 (44)	32.4 (44)	33.8 (53)	34.5 (53)	33.5 (52)	26.6 (50)

NOx Pollution Potentials

STEM-AND-LEAF DISPLAY
LEAF DIGIT UNIT = 1.0000
1 2 REPRESENTS 12.

```
 17   +0*  11122333334444444
(14)  +0·  55677777888899
 29    1*  0111234
 22    1·  5578
 18    2*  113
 15    2·  668
 12    3*  2222
  8    3·  578

      HI    59,   63,   66,  171,  319
```

log (NOx Pollution Potentials)

STEM-AND-LEAF DISPLAY
LEAF DIGIT UNIT = 0.1000
1 2 REPRESENTS 1.2

```
  3   +0*  000
  5   +0T  33
 10   +0F  44444
 20   +0S  6666666667
(11)  +0·  88888999999
 29    1*  000001111
 20    1T  22333
 15    1F  4445555555
  5    1S  77
  3    1·  8
  2    2*
  2    2T  2
  1    2F  5
```

Figure 6. Stem-and-leaf displays of NOx pollution potentials and log (NOx pollution potentials). Note the protection from extraordinary values afforded by stem-and-leaf displays.

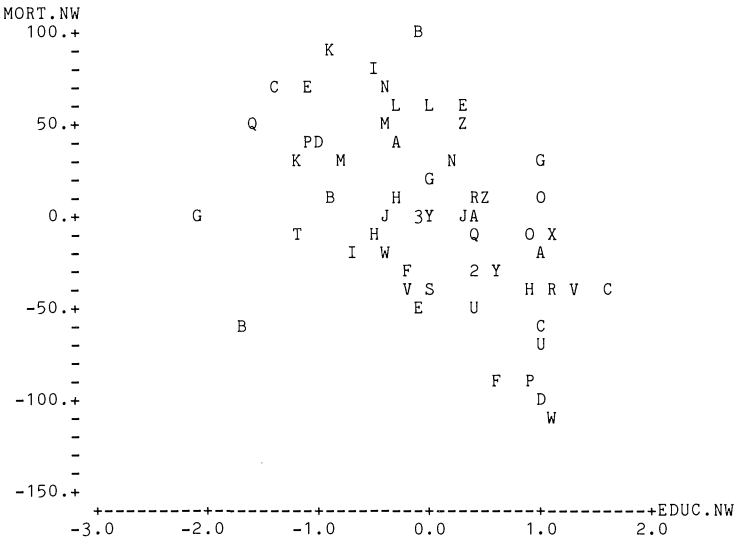

Figure 7. Partial regression plot of MORT against EDUC after removing NONW. Extraordinary values are York (G) and Lancaster (B), Pa.

%NONWHITE, and (partial) correlations of the residuals, MORT · NW, with the remaining candidate carriers, were computed.

Median education (EDUC) was suggested (partial $r^2 = .242$). The partial regression plot for EDUC (Fig. 7) reveals two extraordinary points, Lancaster and York, Pennsylvania, that have unusually low education and low mortality. This may be due to the large Amish populations in the area of these SMSAs, a phenomenon likely to confound rather than assist our understanding of the effects of air pollution on mortality. Because their high leverage influenced the regression away from the trend of the other SMSAs, we set aside York and Lancaster for later examination. (This also eased our qualms about the population density value for York.)

The variables LOGSO2 (partial $r^2 = .190$) and population density, POPDEN, (partial $r^2 = .182$) were now nominated. (Deleting York had enhanced the predictive strength of POPDEN.) Because we hoped to account for all non-pollution variables first, we overruled the 'objective' choice of the higher partial r^2 and selected POPDEN at this step.

The remaining partial correlations were all relatively unexciting on the surface, but some variables showed large partial leverage values, which suggested that the correlations might be deceiving. Figure 8 shows the partial regression plot (removing %NONWHITE, EDUC and POPDEN) for mean January temperature. The linear trend here is largely due to the partial leverage of Miami ($\eta = .226$) and Minneapolis ($\eta = .100$). Three southern California cities, San Diego, Los Angeles and San Jose, form an identifiable group in the lower right with a strong combined influence. We chose not to include January temperature, although it had been included in the previously published analyses, because we would have been devoting a parameter to fitting few data points.

By contrast, the partial regression plot for mean annual precipitation (RAIN) (Fig. 9), shows that Miami ($\eta = .171$) has sharply depressed its predictive value. (Note that these two variables alone account for much of the large leverage exhibited by Miami in the 16-carrier model.) We omitted Miami, and fitted the regression model shown in Table 7.

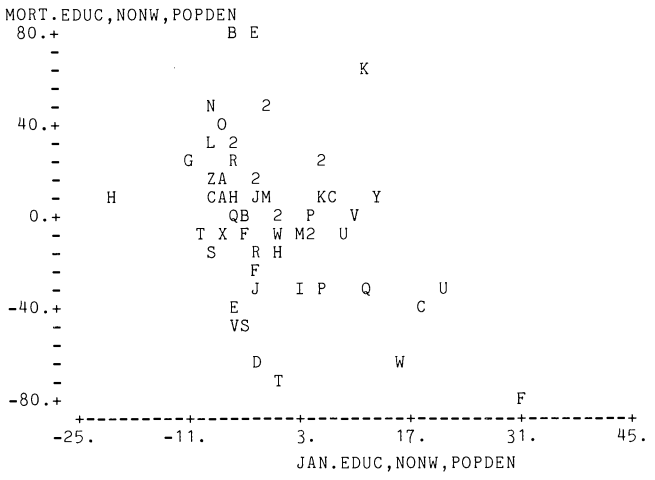

Figure 8. Partial regression plot for mean January temperature removing EDUC, NONW and POPDEN (York and Lancaster omitted). Note extraordinary points: Minneapolis (H), Miami (F), and southern California cities: San Diego (U), Los Angeles (C) and San Jose (W).

No other nonpollution variables improved the model, so we computed and plotted the partial regression plots (removing %NONWHITE, EDUC, POPDEN and RAIN for each of the three air pollution measures). Although LOGNOX shows the largest partial r^2 (.069), the partial regression plot of LOGSO2 (Fig. 10) reveals that New Orleans has depressed its effect ($\eta = .134$). We had no grounds for preferring either of these variables, but they were too closely related to include both. Because they measured related quantities and were measured on equivalent scales, we constructed the combined variable LGNXSO2 = log (NOX + SO2), included it in the model, and omitted New Orleans. We note that a researcher might have reasons for preferring either the model with LGSO2 or that with LGNOX (and New Orleans). All three models fit almost equally well.

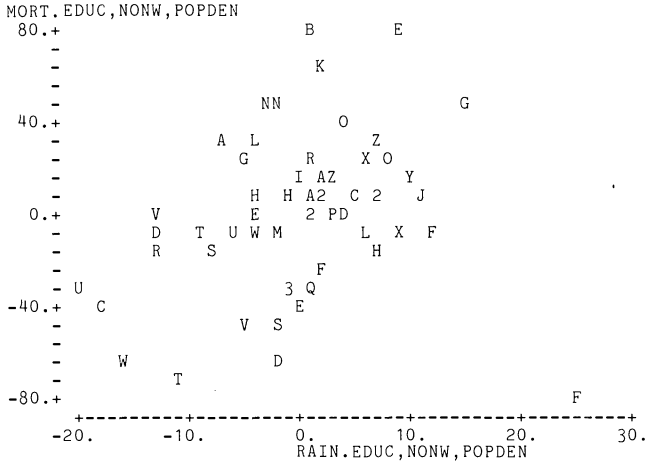

Figure 9. Partial regression plot for mean annual rainfall. Removing EDUC, NONW and POPDEN (York and Lancaster omitted). Note extraordinary point: Miami (F).

Table 7
Regression of MORT on NONW, EDUC, POPDEN and RAIN, omitting York and Lancaster, Pa, and Miami, Fl.

	Column	Coefficient	SD of coefficient	t ratio = coeff./SD
	—	1037.60	83.75	12.39
x_1	NONW	2.709	0.508	5.34
x_2	EDUC	−23.55	6.20	−3.80
x_3	POPDEN	0.01614	0.00330	4.89
x_4	RAIN	1.963	0.543	3.61

SD of y about regression: $s = 30.38$ with $57 - 5 = 52$ df.
$R^2 = 77.3\%$
$R^2 = 75.5\%$, adjusted for df.

	y MORT	Predicted y	SD of pred. y	Residual	Stand. resid.
Albany, N.Y.	997.87	925.85	6.05	72.03	2.42R
Birmingham, Al.	1030.38	1059.42	12.77	−29.04	−1.05X
Nashville, Tn.	1113.16	1051.45	10.96	61.71	2.18R
New Haven, Ct	994.65	1019.14	13.26	−24.49	−0.09X
Toledo, Oh.	967.80	984.27	13.62	−16.47	−0.61X
Utica, N.Y.	823.76	887.09	6.85	−63.33	−2.14R
Washington, D.C.	1003.50	943.49	6.79	60.01	2.03R

R denotes an observation with a large standardized residual.
X denotes an observation whose x value gives it large influence.

The resulting model is shown in Table 8. To permit fair comparisons with other models and to emphasize that the extraordinary points have been treated specially but not forgotten, we have introduced dummy variables for the four extraordinary SMSAs. This 10-carrier regression has an adjusted R^2 of 81.6, and t statistics between 4.89 and 2.86 (excluding the intercept). As Table 6 shows, it appears to fit better than previously

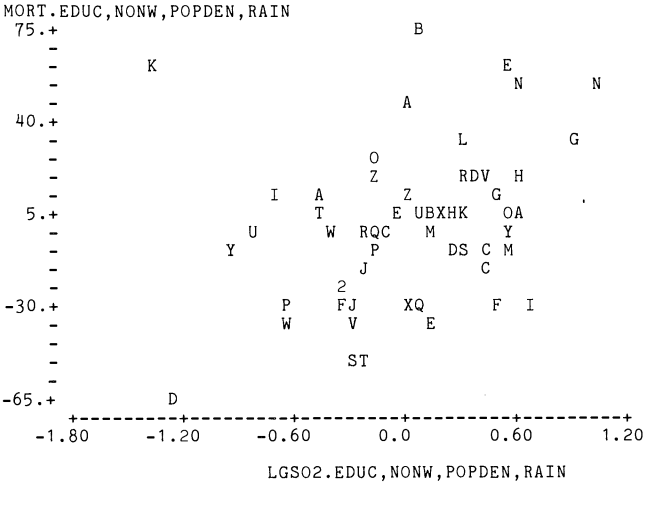

Figure 10. Partial regression plot for LOGSO2 removing EDUC, NONW, POPDEN and RAIN (York, Lancaster and Miami omitted). Note extraordinary point: New Orleans (K).

Table 8
Final regression model for air pollution mortality. Cities requiring special attention have been assigned dummy variables.

	Column	Coefficient	SD of coefficient	t ratio = coeff./SD
	—	926.31	78.92	11.74
x_1	NONW	2.088	0.469	4.45
x_2	EDUC	−16.89	5.68	−2.97
x_3	POPDEN	0.01046	0.00346	3.02
x_4	RAIN	2.467	0.505	4.89
x_5	LGNOXSO2	29.12	8.62	3.38
x_6	YORK	−128.8	34.9	−3.69
x_7	LANCASTER	−113.9	28.5	−4.00
x_8	MIAMI	−104.4	31.1	−3.35
x_9	NEW ORLEANS	82.1	28.8	2.86

SD of y about regression: $s = 26.65$ with $60 - 10 = 50$ df.
$R^2 = 84.4\%$
$R^2 = 81.6\%$, adjusted for df.

	y MORT	Predicted y	SD of pred. y	Residual	Stand. resid.
Albany, N.Y.	997.87	928.22	5.34	69.65	2.67R
Toledo, Oh.	972.46	919.76	5.77	52.70	2.03R

R denotes an observation with a large standardized residual.

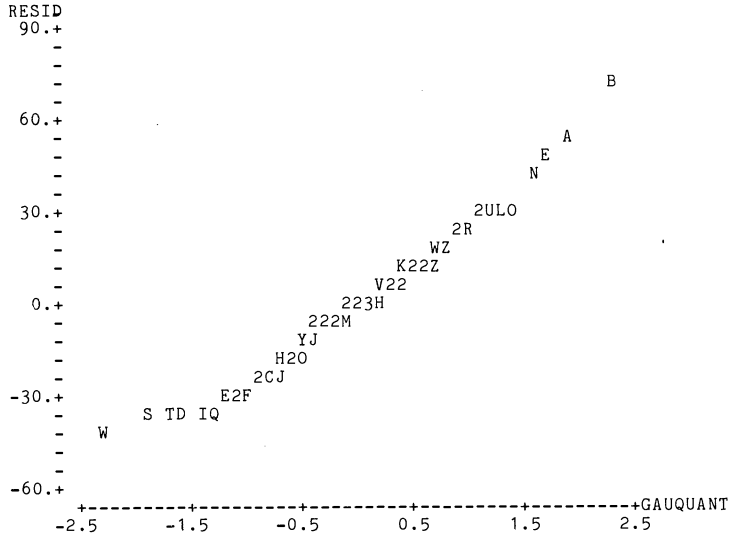

Figure 11. Gaussian quantile–quantile plot of residuals from the regression of Table 7.

proposed models for these data. The largest residual is for Albany, New York (standardized residual = 2.67). The largest leverage values are for Washington, D.C. ($h = .215$) and Birmingham, Alabama ($h = .206$), but neither is particularly extreme. Figure 11 shows the Gaussian Q–Q plot of the residuals.

6. Summary

Inexpensive computing and widely-distributed statistical packages have made blind automated regression model building common. Unfortunately, this approach hides the data from the data analyst and arbitrarily makes decisions that might be better made by the data analyst or the subject-discipline expert. The spread of interactive computing has now made possible a new mode of computer-assisted data analysis: a collaboration between the data analyst and the computer. In such a collaboration, the computer provides a window into the data that enables the data analyst to employ his subject-discipline knowledge in making decisions about the regression model.

We have discussed some elementary techniques to facilitate such analyses, and have illustrated them with reanalyses of two data sets. These examples show some of the patterns that can be revealed by this kind of data analysis. Most of these patterns eluded the automated regression methods employed by previous authors, and several seem to have significantly altered the automated analyses.

It is important to stress that we have not attempted an exhaustive review of proposed or possible techniques for interactive regression model building and that the analyses discussed are illustrative only and not put forward as the 'correct' models for these data. We hope to see additional techniques developed in the coming years.

Acknowledgements

This paper is a longer version of an invited address presented to the Institute of Mathematical Statistics Special Topics Conference on Regression held in Boulder, Colorado in October 1979.

Harold V. Henderson was supported by a New Zealand National Research Advisory Council Research Fellowship during his sojourn at Cornell University. The authors thank R. R. Hocking and G. C. McDonald for making available the data used in the examples.

Résumé

Les techniques automatiques pour la construction de modèles en régression multiple masquent souvent à l'analyste des aspects importants des données analysées. Des caractéristiques comme la non-linéarité, la colinéarité, la présence d'observations suspectes peuvent affecter gravement les analyses automatiques sans être pour autant décelables. Une alternative consiste à utiliser le calcul intéractif et les méthodes exploratoires pour révéler des caractèristiques inattendues dans les données. Un avantage important de cette approche est que l'analyste peut utiliser sa connaissance du sujet pour résoudre les difficultés. Ces méthodes sont illustrées par un retour sur l'analyse de deux ensembles de données utilisées par Hocking (1976, *Biometrics* **32,** 1–44) pour illustrer l'emploi de méthodes automatiques en régression.

References

Belsley, D., Kuh, E. and Welsch, R. E. (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. New York: Wiley.

Chernoff, H. (1973). The use of faces to represent points on k-dimensional space graphically. *Journal of the American Statistical Association* **68,** 361–368.

Cook, R. D. (1977). Detection of influential observations in linear regression. *Technometrics* **19,** 15–18.

Dixon, W. J. and Brown, M. B. (eds) (1979). *BMDP*-79. Biomedical Computer Programs P-Series. Berkeley, California: University of California Press.

Ezekiel, M. (1924). A method of handling curvilinear correlation for any number of variables. *Journal of the American Statistical Association* **19,** 431–453.

Helwig, J. T. and Council, K. A. (eds) (1979). *SAS User's Guide.* Raleigh, North Carolina: SAS Institute, Inc.

Hoaglin, D. C. and Welsch, R. E. (1978). The hat matrix in regression and ANOVA. *American Statistician* **32,** 17–22.

Hocking, R. R. (1976). The analysis and selection of variables in linear regression. *Biometrics* **32,** 1–44.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: application to nonorthogonal problems. *Technometrics* **12,** 55–67.

Larsen, W. A. and McCleary, S. J. (1972). The use of partial residual plots in regression analysis. *Technometrics* **14,** 781–790.

Mallows, C. L. (1973). Some comments on C_p. *Technometrics* **15,** 661–675.

McDonald, G. C. and Schwing, R. C. (1973). Instabilities of regression estimates relating air pollution to mortality. *Technometrics* **15,** 463–481.

McDonald, G. C. and Ayers, J. A. (1978). Some applications of the 'Chernoff faces': A technique for graphically representing multivariate data. In *Graphical Representation of Multivariate Data,* P. C. C. Wang (ed.). New York: Academic Press.

Mosteller, F. and Tukey, J. (1977). *Data Analysis and Regression.* Reading, Massachusetts: Addison–Wesley.

Ryan, T. A., Joiner, B. L. and Ryan B. F. (1976). *Minitab Student Handbook.* North Scituate, Massachusetts: Duxbury Press.

Ryan, T. A., Joiner, B. L. and Ryan, B. F. (1980). *Minitab Manual, Version* 80.1. State College, Pennsylvania: Pennsylvania State University.

Sylvester, J. J. (1884). Lectures on the principles of universal algebra. *American Journal of Mathematics* **6,** 270–286.

Tukey, J. (1977). *Exploratory Data Analysis.* Reading, Massachusetts: Addison–Wesley.

Velleman, P. F. and Hoaglin, D. C. (1981). *Applications, Basics, and Computing of Exploratory Data Analysis.* Boston, Massachusetts: Dxubury Press.

Wilk, M. B. and Gnanadeskian, R. (1964). Probability plotting methods for the analysis of data. *Biometrika* **55,** 1–77.

Wood, F. S. (1973). The use of individual effects and residuals in fitting equations to data. *Technometrics* **15,** 677–686.

Received January 1980; *revised August* 1980