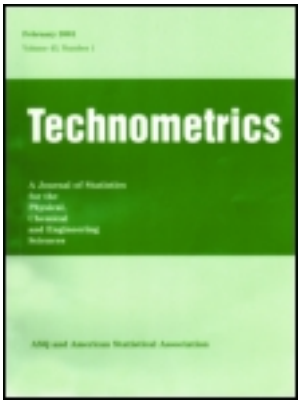

Technometrics
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/utch20

Instabilities of Regression Estimates Relating Air Pollution to Mortality

Gary C. McDonald [a] & Richard C. Schwing [a]
[a] General Motors Research Laboratories, Warren, Michigan
Version of record first published: 09 Apr 2012.

TECHNOMETRICS© VOL. 15, NO. 3 AUGUST 1973

Instabilities of Regression Estimates Relating Air Pollution to Mortality

GARY C. MCDONALD AND RICHARD C. SCHWING

General Motors Research Laboratories
Warren, Michigan

The instability of ordinary least squares estimates of linear regression coefficients is demonstrated for mortality rates regressed around various socioeconomic, weather and pollution variables. A ridge regression technique presented by Hoerl and Kennard (*Technometrics 12* (1970) 69–82) is employed to arrive at "stable" regression coefficients which, in some instances, differ considerably from the ordinary least squares estimates. In addition, two methods of variable elimination are compared—one based on total squared error and the other on a ridge trace analysis.

KEY WORDS

Multiple Linear Regression
Mortality Rate
Pollution Potentials
Ridge Regression
Standardized Total Squared Error, C_p
Ridge Elimination

1. INTRODUCTION

Recently, Lave and Seskin [16] and Hickey, *et al* [9] using multiple regression analyses on large data banks of annual mortality rates and pollution measurements have provided a link between long-term, high pollution levels (sulfates, particulates and heavy metals eminating primarily from stationary sources) and increased mortality. Comparable health studies, either time series or cross section, for the pollutants oxidant, NO_x (oxides of nitrogen) and CO (carbon monoxide) are also available. Hexter and Goldsmith [8], in a recent multiple regression time series study on acute episodes, conclude that carbon monoxide contributes to excess deaths. They did not find an effect due to oxidant. Another study by Shy, *et al* [22] compares illness rates in "clean" versus "polluted" (with NO_x and particulates) neighborhoods in Chattanooga, Tennessee. To eliminate the fact that pollution levels can be correlated with other variables which can influence health, it is often necessary to investigate several highly correlated (non-orthogonal) variables simultaneously. Recently the utility of relatively new statistical techniques for handling such systems has been demonstrated by Hoerl and Kennard [11, 12] and others.

We have chosen to study the chronic effects, as measured by an overall mortality rate, of HC (hydrocarbons), NO_x and SO_2 (sulfur dioxide) employing a "ridge regression" analysis. Because many of the explanatory variables are highly correlated, techniques for estimating the true variable effects for non-orthogonal systems are emphasized. Statistical methods do not in themselves establish a

Received Sept. 1971; revised Aug. 1972

cause-effect link; but assuming the link is present, methods are available to quantify relative contributions of the variable under investigation. In this paper we apply a method termed "ridge regression" to arrive at regression coefficients for the total mortality rate. In addition, for the total mortality rate we eliminate "superfluous" explanatory (or predicting) variables by two methods—one based on total squared error and another on ridge analysis—and compare the results.

2. Description of Variables Considered

A multiple linear additive model will be assumed throughout this paper, i.e., the response variable will be expressed as a linear combination of many explanatory variables. In this section the variables are described and descriptive statistics of the sample used in this study are provided. The total age adjusted mortality rate, our response variable in each regression equation, can be obtained for the years 1959–1961 for 201 Standard Metropolitan Statistical Areas (SMSA) from Duffy and Carroll [4]. In Table 5 of [4], the age-adjusted death rates are given for the categories male white, female white, male non-white and female non-white. In addition, the number of deaths in each of these four categories is also provided. We define our total age adjusted mortality rate to be

$$MR = (\sum D_i)\,(\sum (D_i/R_i))^{-1}, \tag{2.1}$$

where D_i and R_i are the deaths and age adjusted death rates of, say, the ith category respectively, $i = 1, 2, 3, 4$. The sums are then taken over the four categories.

We include three of the explanatory variable groups which are considered important in an epidemiological study of this type. The first fifteen variables listed in Table I can be grouped as follows:

1. Weather
2. Socioeconomic
3. Pollution

Accurate ambient concentration measurements on the air breathed by all residents in an SMSA would be the preferred pollution variables; however, several problems exist with the data which has been published thus far.

(i) Sampling methods and analytical techniques differ between communities.
(ii) Sampling sites are often not representative of the community.
(iii) The distribution of exposures cannot be characterized by a single measure.
(iv) Only eight SMSA's have been studied for the pollutants of primary concern to this study.

Because the above difficulties with available ambient air measurements are so great, we choose to use calculated relative pollution potentials in each SMSA based on emission and weather factors. The pollution potential of three pollutants, namely HC, NO_x , SO_2 , have been estimated by Benedict [1]. The pollution potential is determined as the product of the tons emitted per day per square kilometer of each pollutant and a dispersion factor which accounts for mixing height, wind speed, number of episode days and dimension of each SMSA. Since each SMSA has the same dispersion factor for each pollutant, this quantity is "confounded" with each pollution potential term. Benedict's pollution potentials are available for sixty SMSA's, for the year 1963, which are geographically consistent with the available mortality data. Note, however, that the time period for which the pollution potentials apply (1963) is slightly later than the time period applicable to the mortality data.

Though the pollution variables are labeled HC, NO_x, and SO_2, there are other variables, especially other pollutants, which are highly correlated with each of these indices. For example, SO_2 is highly correlated with certain types of particulates and HC is closely tied to carbon monoxide and lead salts. Thus one cannot demonstrate a specific cause and effect even though the analysis quantifies the relationship.

Previous workers, e.g., Glasser and Greenburg [6], Holland, *et al* [14], and Oechsli and Buechley [20], have found climate or weather variables account for some of the variation in disease rates. Precipitation, mean January temperature, mean July temperature and mean annual humidity, have been included in the present study. These variables, presented in Table I, are considered independently because of their possible effect on health, not because they affect the pollutants.

Several socioeconomic variables are important to account for health differences between communities. Green [7] has suggested indices to optimize the prediction of family health actions from socioeconomic information. Table I includes so-

TABLE I

Description of Variables

Variable Number	Description [Source]
1	Mean annual precipitation in inches, [30].
2	Mean January temperature in degrees Fahrenheit, [30].
3	Mean July temperature in degrees Fahrenheit, [30].
4	Percent of 1960 SMSA population which is 65 years of age or over, [5].
5	Population per household, 1960 SMSA, [24,25].
6	Median school years completed for those over 25 in 1960 SMSA, [27].
7	Percent of housing units which are sound with all facilities, [24].
8	Population per square mile in urbanized area in 1960, [23,25].
9	Percent of 1960 urbanized area population which is non-white, [26].
10	Percent employment in white-collar occupations in 1960 urbanized area, [28].
11	Percent of families with income under $3,000 in 1960 urbanized area, [28].
12	Relative pollution potential of hydrocarbons, HC, [1].
13	Relative pollution potential of oxides of nitrogen, NO_x, [1].
14	Relative pollution potential of sulfur dioxide, SO_2, [1].
15	Percent relative humidity, annual average at 1 p.m., [29].
16	Total age adjusted mortality rate, all causes [4] and equation (2.1), expressed as deaths per 100,000 population.

cioeconomic terms which account for broad differences in occupation, population density, education, income, housing, race and age.

Table II gives the sample means, standard deviations, and the minimum and maximum values for each of our variables. An examination of the data indicates a "bunching" of the pollution potential variables at values below their means. This is the result of including in our analysis several SMSA's which have relatively high values of the pollution potential variables. In particular, Los Angeles and San Francisco have the two largest HC pollution potential values, 648 and 311 respectively, while the third largest value is 144. Los Angeles and San Francisco also have the two largest NO_x values, 319 and 171 respectively, while the third largest value is 66. The SO_2 variable is more evenly distributed among our sample of sixty SMSA's. Table III provides the correlations among the variables considered. The largest correlation, .9838, occurs between the HC and NO_x pollution potentials; other large positive correlations exist between education and percent white collar, and between percent non-white and percent under $3000.

3. A Description of Ridge Regression

Multiple linear regression techniques have played a prominent role in the studies of associations between air pollution and mortality (and/or morbidity) rates. This technique may provide an adequate basis for overall prediction, but, when the explanatory variables are non-orthogonal, it frequently fails to give proper weight to the individual explanatory variables used as predictors. In many problems where data are not obtained from a well designed or controlled experiment, as is the case in air pollution studies involving socioeconomic, weather and other uncontrolled variables, non-orthogonality requires that estimation of individual effects be handled by techniques other than ordinary least squares solutions. Reinke [21] pointed out these difficulties in air pollution models several

Table II

Means, Standard Deviations, Minimum and Maximum Values of Variables (60 Observations)

Variable	Mean	Std. Dev.	Minimum	Maximum
Precipitation	37.37	9.98	10.00	60.00
January Temperature	33.98	10.17	12.00	67.00
July Temperature	74.58	4.76	63.00	85.00
% 65 Years & Older	8.80	1.46	5.60	11.80
Population/Household	3.26	0.14	2.92	3.53
Education	10.97	0.85	9.00	12.30
% Sound Housing	80.92	5.15	66.80	90.70
Population/Mile^2	3876.05	1454.10	1441.00	9699.00
% Non-White	11.87	8.92	0.80	38.50
% White Collar	46.08	4.61	33.80	59.70
% Under $3000	14.37	4.16	9.40	26.40
HC Potential	37.85	91.98	1.00	648.00
NO_x Potential	22.65	46.33	1.00	319.00
SO_2 Potential	53.77	63.39	1.00	278.00
Relative Humidity	57.67	5.37	38.00	73.00
Total Mortality	940.36	62.21	790.73	1113.20

TABLE III

Correlations Among the Variables Considered (60 Observations)

Table III. Correlations Among the Variables Considered (60 Observations)

	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	.0922	.5033	.1011	.2634	-.4904	-.4903	-.0035	.4132	-.2973	.5066	-.5318	-.4873	-.1069	-.0773	.5095
2		.3463	-.3981	-.2092	.1163	.0139	-.1001	.4538	.2380	.5653	.3508	.3210	-.1078	.0679	-.0300
3			-.4340	.2623	-.2385	-.4155	-.0610	.5753	-.0214	.6193	-.3565	-.3377	-.0993	-.4528	.2770
4				-.5091	-.1389	.0649	.1620	-.6378	-.1177	-.3098	-.0205	-.0021	.0172	.1124	-.1746
5					-.3951	-.4095	-.1843	.4194	-.4257	.2599	-.3882	-.3584	-.0041	-.1357	.3573
6						.5515	-.2439	-.2088	.7032	-.4033	.2868	.2244	-.2343	.1765	-.5110
7							.1806	-.4091	.3376	-.6806	.3859	.3476	.1180	.1224	-.4248
8								-.0057	-.0318	-.1629	.1203	.1653	.4321	-.1250	.2655
9									-.0044	.7049	-.0259	.0184	.1593	-.1180	.6437
10										-.1852	.2037	.1600	-.0685	.0607	-.2848
11											-.1298	-.1025	-.0965	-.1522	.4105
12												.9838	.2823	-.0202	-.1772
13													.4094	-.0459	-.0774
14														-.1026	.4259
15															-.0885

years ago, and suggested that ridge analysis, as described by Hoerl [10] and Draper [3], provides a promising method for avoiding distortions such as described above. The recent papers of Hoerl and Kennard [11, 12] and Marquardt [18] give an excellent description of the theory and applications of what has now been termed "ridge regression."

As has been shown in [11], the estimates of regression coefficients tend to become too large in absolute values and it is possible that some will even have the wrong sign. The chances of encountering such difficulties increase the more the prediction vectors deviate from orthogonality. Consider the standard model for multiple linear regression,

$$\mathbf{y} = \mathbf{x}\beta + \varepsilon, \tag{3.1}$$

where $E(\varepsilon) = \mathbf{0}$, $E(\varepsilon\varepsilon') = \sigma^2 \mathbf{I}_n$ and $\mathbf{x}$ is $(n \times p)$ and of full rank. The variables are assumed to be standardized so that $\mathbf{x}'\mathbf{x}$ is in the form of a correlation matrix, and the vector $\mathbf{x}'\mathbf{y}$ is the vector of correlation coefficients of the response variable with each of the explanatory variables. Let

$$\hat{\beta} = (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{y} \tag{3.2}$$

be the least squares estimate of β. The difficulties in this standard estimation are a direct consequence of the average distance between $\hat{\beta}$ and β. In particular, if L^2 is the squared distance between $\hat{\beta}$ and β, then the following hold:

$$\begin{aligned} L^2 &= (\hat{\beta} - \beta)'(\hat{\beta} - \beta), \\ E(L^2) &= \sigma^2 \operatorname{trace} (\mathbf{x}'\mathbf{x})^{-1}, \\ E(\hat{\beta}'\hat{\beta}) &= \beta'\beta + \sigma^2 \operatorname{trace} (\mathbf{x}'\mathbf{x})^{-1}. \end{aligned} \tag{3.3}$$

In terms of the eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0$ of $\mathbf{x}'\mathbf{x}$, we can write

$$\begin{aligned} E(L^2) &= \sigma^2 \sum_{i=1}^{p} \lambda_i^{-1} > \sigma^2 \lambda_p^{-1}, \\ E(\hat{\beta}'\hat{\beta}) &= \beta'\beta + \sigma^2 \sum_{i=1}^{p} \lambda_i^{-1} > \beta'\beta + \sigma^2 \lambda_p^{-1}, \end{aligned} \tag{3.4}$$

and when the error is normally distributed

$$\text{Var}(L^2) = 2\sigma^4 \sum_{i=1}^{p} \lambda_i^{-2} > 2\sigma^4 \lambda_p^{-2}. \tag{3.5}$$

As the vectors of $\mathbf{x}$ deviate further from orthogonality, λ_p becomes smaller and $\hat{\beta}$ can be expected to be farther from the true parameter vector β.

Ridge regression is an estimation procedure based upon

$$\hat{\beta}^* \equiv \hat{\beta}^*(k) = [\mathbf{x}'\mathbf{x} + k\mathbf{I}]^{-1}\mathbf{x}'\mathbf{y}, \quad k \geq 0, \tag{3.6}$$

and has two aspects. The first is the ridge trace which is a two-dimensional plot of the $\hat{\beta}_i^*(k)$ and the residual sum of squares, $\varphi^*(k)$, for values of k in the interval [0, 1]. The trace serves to portray the complex interrelationships that exist between non-orthogonal prediction vectors and the effect of these interrelationships on the estimation of β. The second aspect is the determination of a value of k that gives a better estimate of β by dampening the effect of (3.4). It should be noted that the estimators $\hat{\beta}^*(k)$ are biased estimators when $k > 0$; of course, at $k = 0$ these estimators reduce to those of ordinary least squares which are unbiased.

The vector $\hat{\beta}^*$ for $k > 0$ is shorter than $\hat{\beta}$, i.e., $(\hat{\beta}^*)'(\hat{\beta}^*) < \hat{\beta}'\hat{\beta}$. In fact, $(\hat{\beta}^*)'(\hat{\beta}^*)$ is a decreasing function in $k > 0$. For an estimate $\hat{\beta}^*$ the residual sum of squares is given by

$$\begin{aligned} \varphi^* \equiv \varphi^*(k) &= (\mathbf{y} - \mathbf{x}\hat{\beta}^*)'(\mathbf{y} - \mathbf{x}\hat{\beta}^*) \\ &= \mathbf{y}'\mathbf{y} - (\hat{\beta}^*)'\mathbf{x}'\mathbf{y} - k(\hat{\beta}^*)'(\hat{\beta}^*). \end{aligned} \tag{3.7}$$

The $\mathbf{y}'\mathbf{y}$ term is the sum of squares of the dependent variable and is equal to 1 when the data are transformed as indicated in this section; the $(\hat{\beta}^*)'\mathbf{x}'\mathbf{y}'$ is the sum of squares due to regression; and $k(\hat{\beta}^*)'(\hat{\beta}^*)$ is an adjustment term associated with the ridge analysis. The coefficient of determination is given by

$$\begin{aligned} R^2 \equiv R^2(k) &= (\hat{\beta}^*)'\mathbf{x}'\mathbf{x}(\hat{\beta}^*) \\ &= (\hat{\beta}^*)'\mathbf{x}'\mathbf{y} - k(\hat{\beta}^*)'(\hat{\beta}^*). \end{aligned} \tag{3.8}$$

Where $\mathbf{x}'\mathbf{x} = I$, i.e., the explanatory variables are uncorrelated, then $\hat{\beta}^*(k) = (k + 1)^{-1}\mathbf{x}'\mathbf{y} = (k + 1)^{-1}\hat{\beta}$. In other words, the least squares coefficients are uniformly scaled by the quantity $(k + 1)^{-1}$. The relative values of the regression coefficients are then independent of the choice of k; i.e., $\hat{\beta}_i^*(k)/\hat{\beta}_j^*(k) = \hat{\beta}_i/\hat{\beta}_j$, $1 \leq i, j \leq p$, $\hat{\beta}_i \neq 0$, for all $k \geq 0$.

4. A Ridge Regression Example: Total Mortality Rate

The $\mathbf{x}'\mathbf{x}$ and $\mathbf{x}'\mathbf{y}$ in correlation form are given in Table III. These values are based on a total of 60 observations. There are several large interfactor correlations, the most notable being that between the pollutant potentials of hydrocarbon and oxides of nitrogen. This is also reflected in the eigenvalues of $\mathbf{x}'\mathbf{x}$ which are:

$\lambda_1 = 4.5272$	$\lambda_6 = .9605$	$\lambda_{11} = .1665$
$\lambda_2 = 2.7547$	$\lambda_7 = .6124$	$\lambda_{12} = .1275$
$\lambda_3 = 2.0545$	$\lambda_8 = .4729$	$\lambda_{13} = .1142$
$\lambda_4 = 1.3487$	$\lambda_9 = .3708$	$\lambda_{14} = .0460$
$\lambda_5 = 1.2227$	$\lambda_{10} = .2163$	$\lambda_{15} = .0049$

The sum of the reciprocals of the eigenvalues is $\sum \lambda_i^{-1} = 263.06$. Thus, from

equation (3.4), the expected squared distance of the least squares coefficient estimate, $\hat{\beta}$, from β is 263.06 σ^2, which is more than seventeen times what it would be for an orthogonal system.

Figure 1 is the ridge trace for this problem. This trace was constructed by computing a total of 21 regressions using $\hat{\beta}^* = [\mathbf{x}'\mathbf{x} + k\mathbf{I}]^{-1}\mathbf{x}'\mathbf{y}$ and 21 equally spaced values of k in the interval [0, 1]. The ridge trace gives a two-dimensional portrayal of the effects of the correlations among the explanatory variables and

RIDGE TRACE: TOTAL MORTALITY (Variables 1–15)

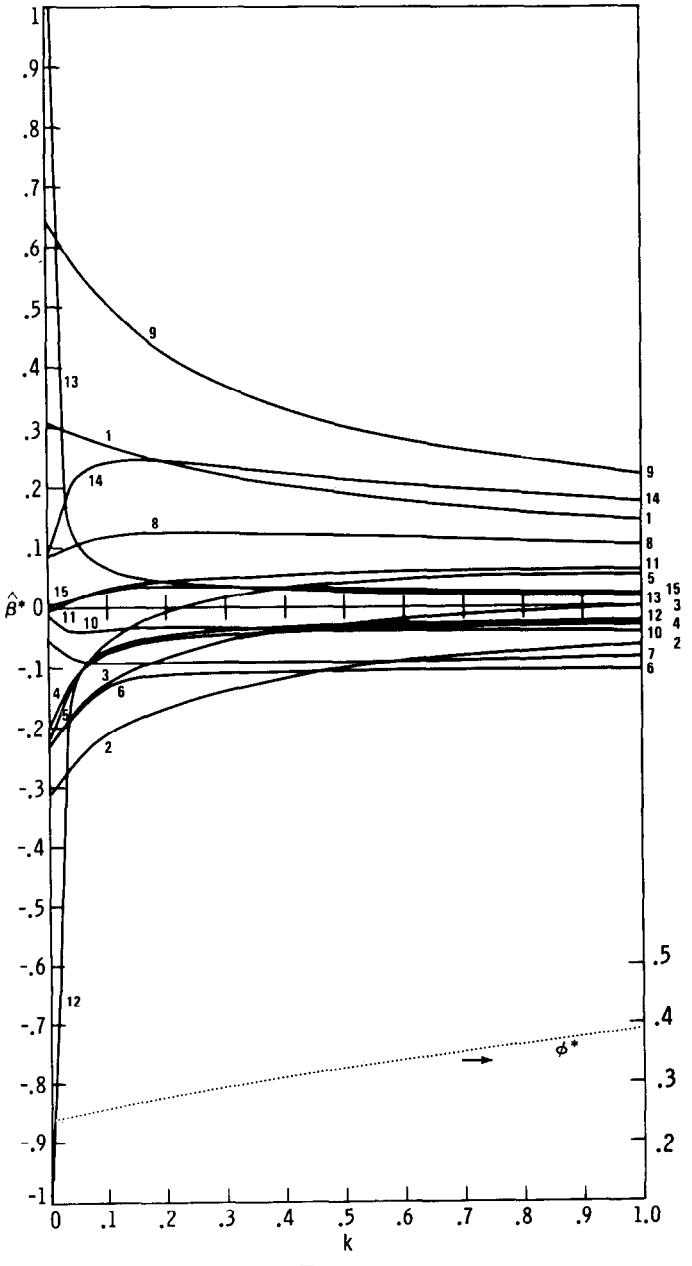

FIGURE 1

makes possible assessments that are usually not made even if all possible regressions are computed. For example:

(i) The coefficients from the ordinary least squares are very likely to be overestimated. At least, they are collectively not stable. Moving a short distance from the least squares point $k = 0$ shows a rapid decrease in absolute value of at least two variables, namely, variables 12 and 13 which are the hydrocarbon and oxides of nitrogen variables respectively. This is not unexpected because these two quantities have a sample correlation coefficient of .98. Both of these coefficients are quickly driven towards zero and are almost mirror images of each other about the zero line.

(ii) The effect of variable 14, the sulfur dioxide term, is likely to be originally underestimated. The coefficient increases as k increases while the magnitudes of the coefficients of the other two pollution variables decrease.

(iii) The effects of variables 2, the mean January temperature, and 9, percent non-white population, also appear to be overestimated in absolute value. Both variables decrease in absolute value as k increases, and level off at non-zero values.

(iv) Variable number 8, the population density factor, is quite stable; the coefficient of this variable moves very little as k ranges between 0 and 1.

(v) The coefficients, with the exception of that corresponding to variable number 5, appear to stabilize in the neighborhood of $k = .2$. We would expect coefficients chosen at this point to be closer to β and more suitable for estimation of individual effects than the least squares coefficients. The residual sum of squares at $k = .2$ has increased about 17% from the corresponding value at $k = 0$.

(vi) The squared length of the coefficient vector decreases rapidly as k increases as shown in Figure 2. At $k = .05$, it is only 23% of its original value; whereas if the least squares coefficients were computed from an orthogonal system, i.e., $\mathbf{x}'\mathbf{x} = I$, it would be 91% of its original value.

5. Elimination of Variables

Total Squared Error

In order to adequately represent the mortality rate data as a linear function in fewer than fifteen explanatory variables, it is essential that some simple criterion of goodness of fit be chosen to characterize each equation. The measure recommended by Daniel and Wood [2] is the "standardized total squared error" given by Mallows [17]. This statistic, called C_p (p is the number of variables in the regression including a constant term if needed), estimates the sum of the squared biases plus the squared random errors in the response variable at all n data points. It is a simple function of the residual sum of squares from each fitting equation. Mallows has shown that regressions with small bias will have C_p's nearly equal to p and so this, as well as the magnitude of C_p , is used to judge a particular subset regression. The quantity C_p is given by

$$C_p = (RSS_p/\hat{\sigma}^2) - (n - 2p), \tag{5.1}$$

where p is the number of variables in the regression, RSS_p is the residual sum of squares for the particular p-variate regression being considered and $\hat{\sigma}^2$ is an estimate of σ^2, the variance of the error term in the regression model. Frequently $\hat{\sigma}^2$ is taken to be the residual mean square from the complete regression.

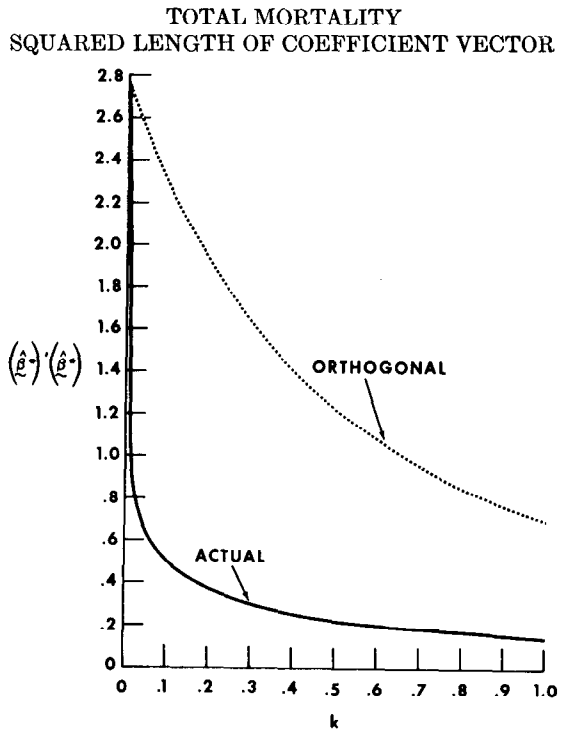

FIGURE 2

If the total number of input variables is not large (say $\leq$ 12) then the residual sum of squares can be computed for each regression equation and compared via a graph of C_p versus p. However, it is not always practical or feasible to compute all possible regressions. Procedures for determining regressions which have small C_p values for each allowable value of p, and in fact to determine which regression minimizes C_p , without actually computing all regressions, are given by Hocking and Leslie [13] and by La Motte and Hocking [15]. We have applied the algorithm and computer program described in [15], to arrive at a "best" regression in the sense of minimizing C_p and to isolate subsets of the fifteen explanatory variables which yield "almost best" regressions. To arrive at the "best" equation using the above method necessitated computation of 1,465 sets of regression estimates, which is about 4.5% of the 32,768 total possible regressions.

Figure 3 is a C_p graph using total mortality rate as the response variable. In our computations, we used as input the raw data, i.e., our regression model for this part of the analysis was not standardized, and so a constant term is counted as a parameter. Thus p may take values up to 16, and equals 16 when all explanatory variables are included. It then follows that $C_p \geq 2p - 16$ for all $p = 1, \cdots, 16$. The regression equation with variables 1, 2, 3, 6, 9 and 14 as the explanatory variables yields the overall minimum value of C_p which is 3.55. This equation can be written as

$$\begin{aligned}\text{Mortality Rate} = {} & 1180.4 + 1.797\ (\text{Precipitation}) - 1.484\ (\text{Jan. Temp.}) \\ & - 2.355\ (\text{July Temp.}) - 13.619\ (\text{Education}) \\ & + 4.585\ (\%\ \text{Non-White}) + 0.260\ (SO_2\ \text{Potential}) + \epsilon. \qquad (5.2)\end{aligned}$$

The coefficient of determination (R^2) for this equation is 0.735. The corresponding

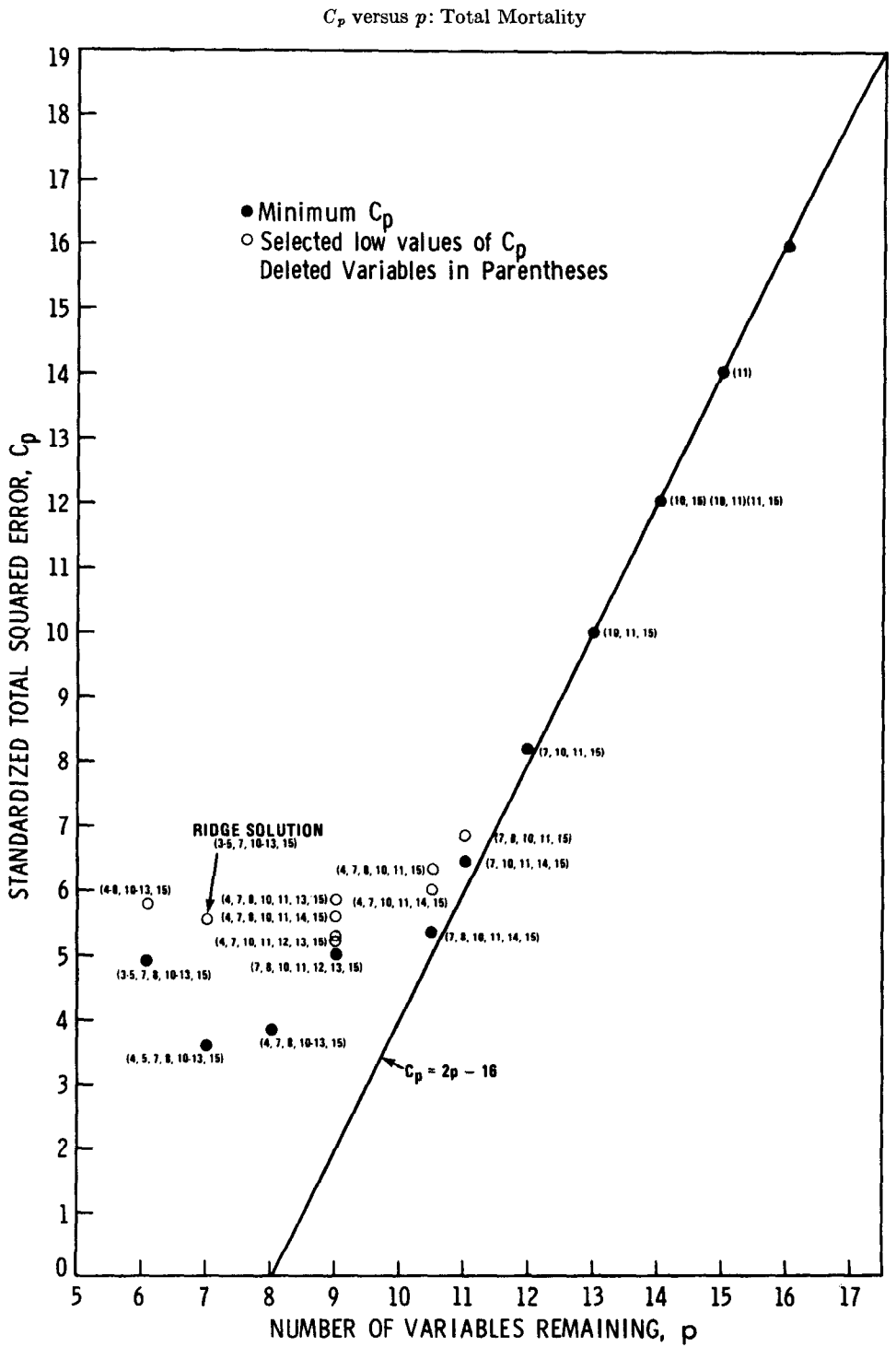

FIGURE 3

value for the full regression with all variables entered is 0.764. Other "almost best" subsets of variables are given in Figure 3. Variables 1, 2, 9 and 14 are contained in almost all of the subsets with small C_p values. The "best" set of five variables is 1, 2, 6, 9 and 14 which yields $C_p = 4.90$.

Elimination Using the Ridge Trace

Hoerl and Kennard [12] suggest a method of variable elimination which is based on the ridge trace. Their procedure is:

(i) Examine the stable coefficients and eliminate the factors with the least predicting power. From our ridge trace in Figure 1, variables which appear stable with coefficients small in absolute value are 4, 7, 10, 11 and 15; hence eliminate these variables.

(ii) Examine the unstable coefficients and eliminate those factors that cannot hold their predicting power. It is obvious from the trace that variables 12 and 13 can be eliminated with this criterion.

(iii) Delete one or more of the remaining unstable coefficients. In our example variables 3 and 5 are being eliminated at this step.

Based on this ridge analysis, the explanatory variables now remaining are 1, 2, 6, 8, 9, 14 which agrees with the subset of variables chosen using the total squared error measure with the exception that variable 3, the mean July temperature, has been replaced with variable 8, the population density. The C_p value associated with this particular subset of variables is 5.52 which is about 55% larger than the minimum value, but closer to the value $p = 7$. The regression (or least squares) equation with variables 1, 2, 6, 8, 9, 14 as the explanatory variables can be written as

$$\begin{aligned}\text{Mortality Rate} = {} & 988.4 + 1.487\ (\text{Precipitation}) - 1.633\ (\text{Jan. Temp.}) \\ & - 11.533\ (\text{Education}) + 0.004\ (\text{Pop. Density}) \\ & + 4.145\ (\%\ \text{Non-White}) + 0.245\ (SO_2\ \text{Potential}) + \epsilon. \qquad (5.3)\end{aligned}$$

The coefficient of determination for this equation is 0.724.

Let $S_1 = \{1, 2, 3, 6, 9, 14\}$ be the subset of variables determined by the minimum C_p criterion and $S_2 = \{1, 2, 6, 8, 9, 14\}$ be that subset obtained by eliminating variables based on the ridge trace. We can now examine the stability of these two subsets of variables via the technique of ridge regression as was done with all fifteen explanatory variables. Figures 4 and 5 are the ridge traces for total mortality rate versus the set of variables specified by S_1 and S_2 respectively. The two traces are quite similar with respect to variables 1, 2, 6, 9 and 14. This collection of variables yield the minimum C_p value for five explanatory variables in addition to a constant term (i.e., $p = 6$). However, variable 3 is quite unstable when compared to the corresponding plot of variable 8 whose value is practically constant while k ranges from 0 to 1. This is also reflected in the sum of the reciprocals of the eigenvalues of the $\mathbf{x}'\mathbf{x}$ matrix in the two cases. With respect to S_1 this sum is 9.86 and for S_2 it is 9.09, about fifty percent larger than the corresponding value for an orthogonal system. Several characteristics of the coefficients, which are common to both of these ridge traces, are worth noting. The effects of precipitation, the percent non-white population and SO_2 potential are increasing, i.e., the coefficients are positive. The precipitation and SO_2 potential are about equal in their relative contribution but substantially below the contribution of the percent non-white factor. The mean January temperature and education terms have a decreasing effect on the mortality rate, but their relative order of importance reverses in the neighborhood of $k = .15$. The education coefficient is quite stable in the given range of k. The relative effect of the mean July temperature is questionable; the sign of the coefficient changes as indicated in the ridge trace, and in terms of absolute value its contribution would certainly be less

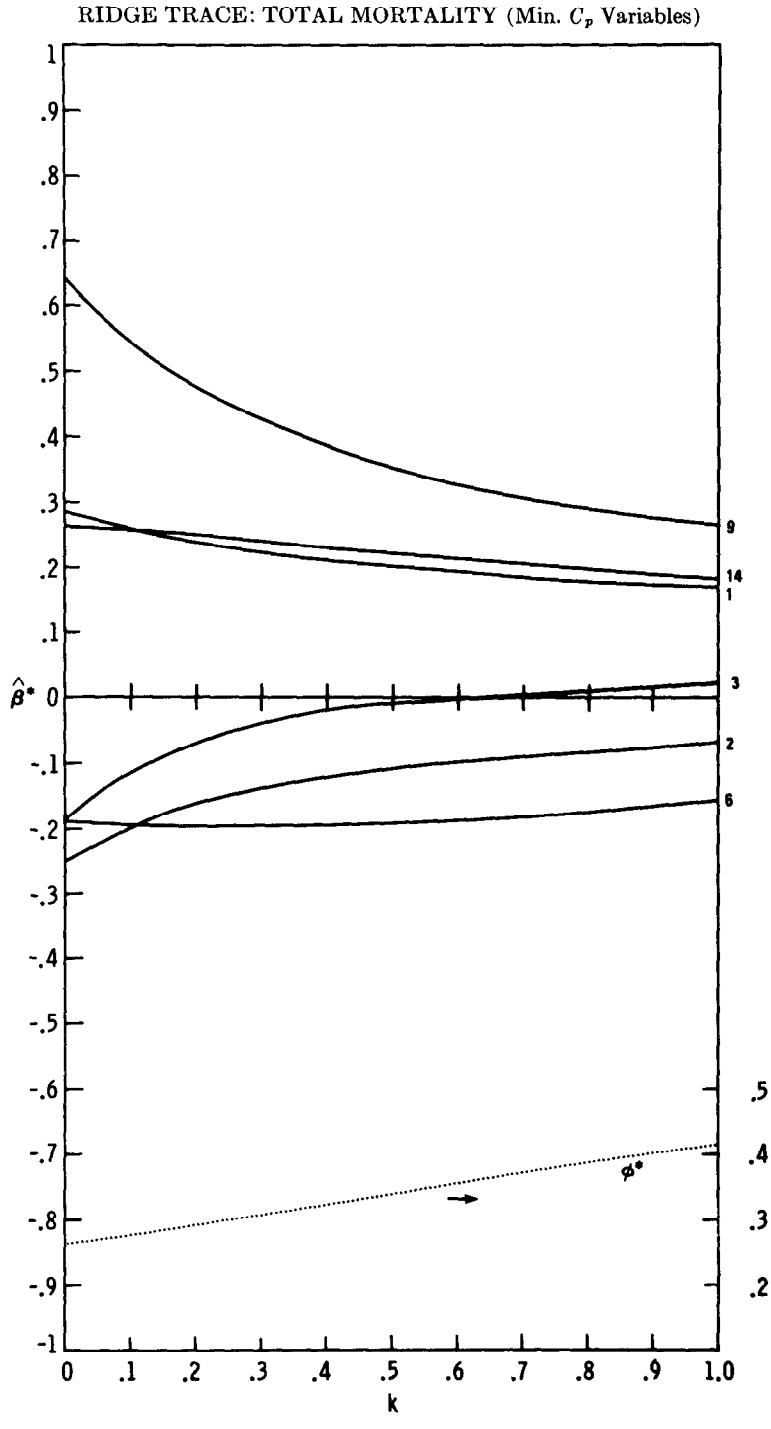

FIGURE 4

than the other variables being considered. Using the C_p criterion this variable would be eliminated in arriving at a "best" subset with five explanatory variables. As noted before, population density has a stable increasing effect.

Figures 6 and 7 are the squared length of the coefficient vectors of the variables

RIDGE TRACE: TOTAL MORTALITY (Ridge Elimination Variables)

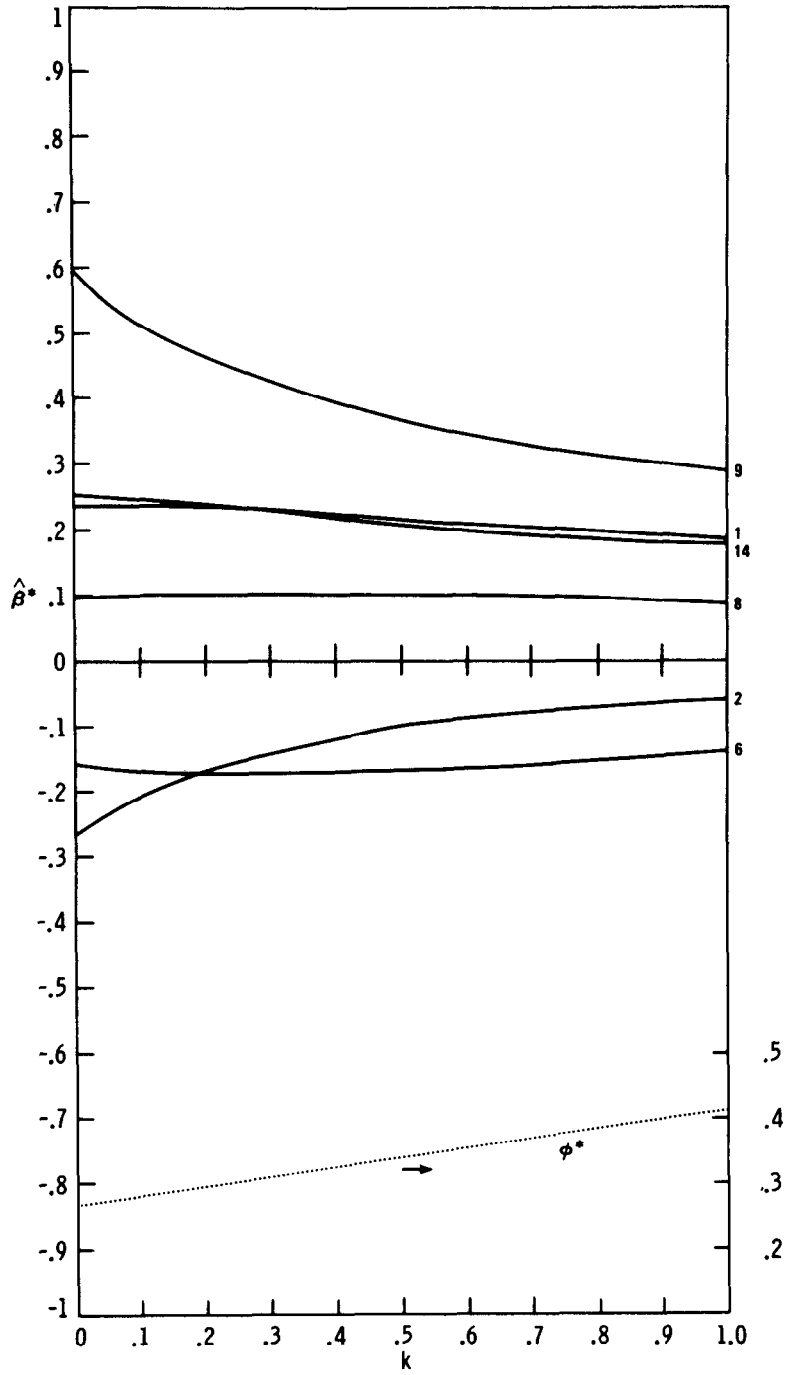

FIGURE 5

specified by S_1 and S_2 respectively. For the S_2 variables, the system does not act unlike an orthogonal system. The decrease in the length of the coefficient vector for this reduced set of variables is almost identical to what it would be in the case of orthogonality.

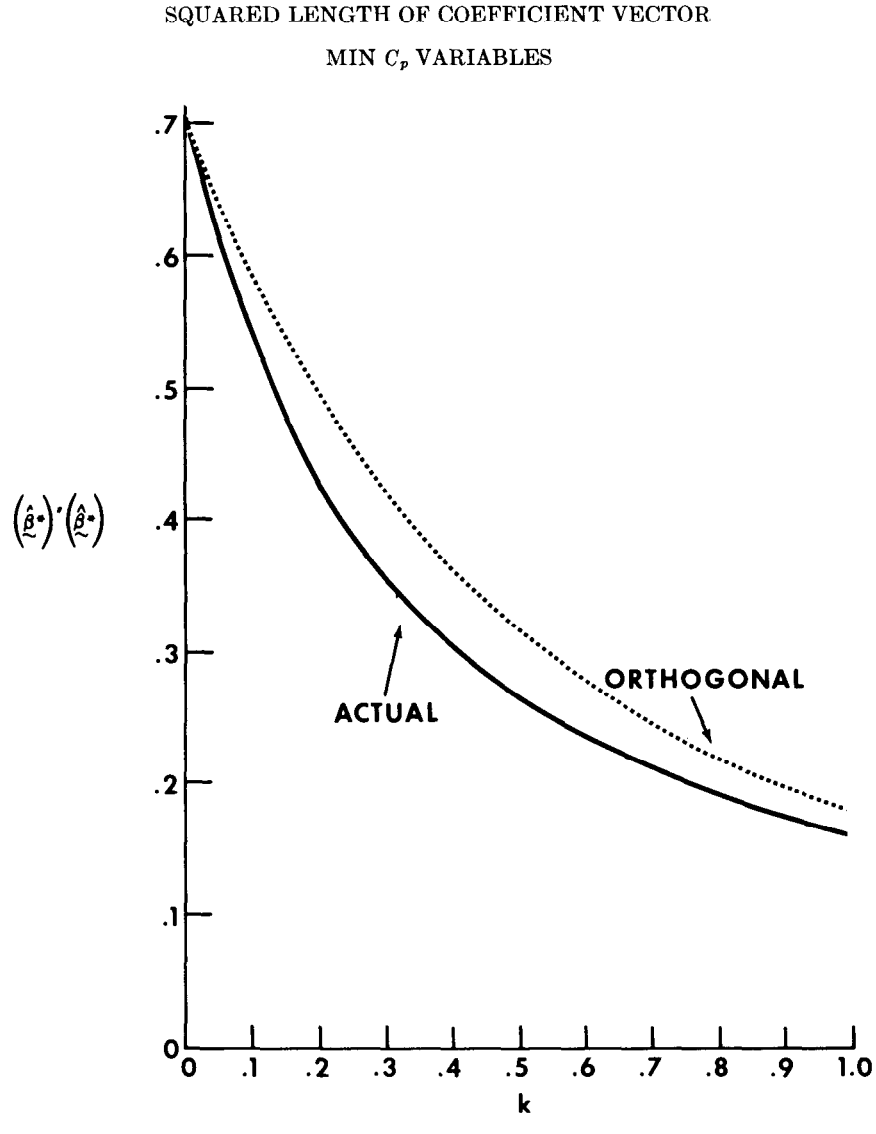

FIGURE 6

6. CONCLUSIONS AND SUMMARY

We have discussed the total mortality data in detail with respect to ridge regression and eliminating explanatory variables employing two different criteria. Table IV provides a summary of our results contrasting the ordinary least squares estimates with those obtained from a ridge trace at the value $k = .2$. This table exemplifies the instability of least squares estimates in this problem—namely, the coefficients which achieve a residual sum of squares slightly larger than the minimum value can differ by more than an order of magnitude, and even in sign, from the corresponding least squares estimates. Entries are given for the regression equation using all fifteen explanatory variables, the six variables selected by the C_p criterion, and the six chosen by the ridge elimination method. The standard deviations of the coefficient estimates are given below the estimates and are enclosed in parentheses. The sum of squares of the estimated coefficients, the

residual sum of squares and the coefficient of determination are provided for each of the derived equations. All coefficients apply to the standardized model described in Section 3. In our analysis, emphasis was placed on a technique which did not eliminate variables, since we were specifically interested in the effects of all the pollution potential variables. Hoerl and Kennard [12] suggest the best strategy is retaining all variables in the analysis and choosing a "good" value of k. Variables with small effects will then have small coefficients. As can be noted, the coefficients of the variables remaining after elimination (by either method) agree rather closely with the corresponding values with all variables included at $k = .2$. This agreement is not as good at $k = 0$, i.e., when considering least squares solutions.

Our choice of the value $k = .2$ is reasonable in the sense that: (i) all major changing of order in the coefficient estimates has already occurred, (ii) the residual sum of squares and coefficient of determination have values consistent with problems of this type, and (iii) assuming normally distributed errors, the vector $\hat{\beta}^*(.2)$ lies interior to the 95% confidence ellipsoid for the unknown vector β. However, this particular value of k is not known to be "optimal" in the sense of minimizing

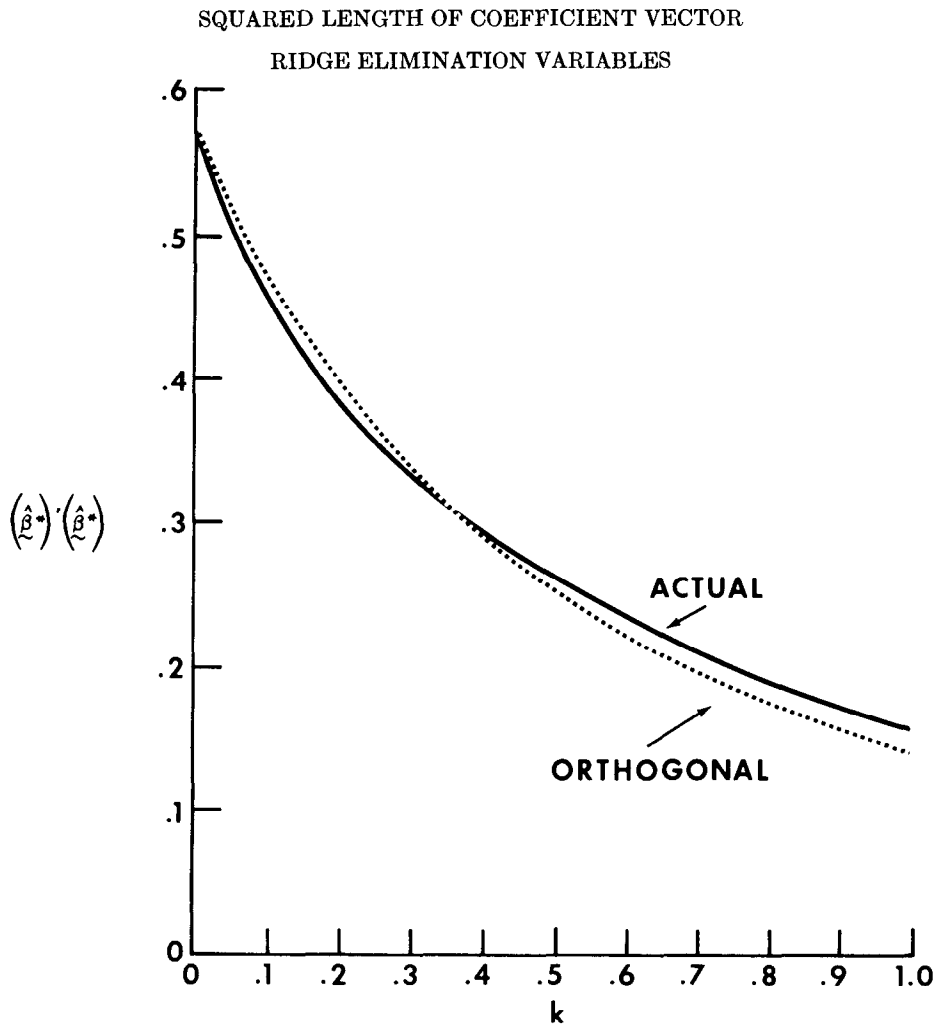

FIGURE 7

TABLE IV

Total Mortality: Coefficients of Standardized Variables for Two Ridge Solutions; $k = 0$ (the Least Squares Solution) and $k = .2$.

The standard deviation of an estimate is given directly below it and in parentheses. Summary statistics are given at the bottom of the table.

	15 Variables		Minimum C_p		Ridge Elimination	
Variable Name	k=0	k=.2	k=0	k=.2	k=0	k=.2
Precipitation	.306	.243	.288	.247	.239	.230
	(.148)	(.069)	(.096)	(.065)	(.094)	(.065)
January Temperature	-.318	-.168	-.242	-.164	-.267	-.172
	(.181)	(.065)	(.084)	(.063)	(.085)	(.063)
July Temperature	-.237	-.084	-.180	-.073		
	(.146)	(.071)	(.095)	(.066)		
% 65 Years and Older	-.213	-.055				
	(.200)	(.063)				
Population per Household	-.232	-.007				
	(.152)	(.068)				
Education	-.233	-.114	-.185	-.190	-.157	-.171
	(.161)	(.070)	(.087)	(.063)	(.090)	(.065)
% Sound Housing Units	-.052	-.094				
	(.146)	(.069)				
Population per Mile[a]	.084	.123			.097	.091
	(.094)	(.065)			(.082)	(.063)
% Non-White	.640	.423	.657	.481	.594	.462
	(.190)	(.068)	(.100)	(.066)	(.094)	(.065)
% White Collar	-.014	-.034				
	(.123)	(.068)				
% Under $3000	-.009	.044				
	(.216)	(.066)				
HC Pollution Potential	-.979	-.046				
	(.724)	(.045)				
NO_x Pollution Potential	.983	.043				
	(.747)	(.043)				
SO_2 Pollution Potential	.090	.243	.265	.255	.249	.232
	(.150)	(.066)	(.080)	(.061)	(.087)	(.064)
Relative Humidity	.009	.033				
	(.101)	(.063)				
$(\hat{\beta}^*)'(\hat{\beta}^*)$	2.758	.380	.711	.426	.578	.388
φ^*	.236	.276	.265	.289	.276	.292
$\bar{R}^2$	.764	.572	.735	.541	.724	.553

the expected squared distance between $\hat{\beta}^*(k)$ and β (or at least to have this distance less than the corresponding distance between $\hat{\beta}^*(0)$ and β). Hoerl and Kennard [11] have established the *existence* of a ridge estimator (i.e., a k-value) which achieves a smaller expected squared distance than the least squares estimator. Newhouse and Oman [19] propose several methods for choosing a k value to use in ridge regression and investigate their properties using Monte Carlo experiments with two explanatory variables. It appears that an optimal choice of k (or interval

of k values) is an open question at this time unless one has prior knowledge about the length and/or direction of the unknown coefficient vector.

We have done similar analyses with the three pollution potential variables replaced in the linear model by the natural logarithms of the pollution potentials. The instability of the least squares estimates is still present using these transformed variables. The values of the coefficients of the three pollution variables at $k = 0$ and $k = .2$ respectively are: (i) ln (HC), $-.669$ and .003; (ii) ln (NO_x), 1.028 and .206; (iii) ln (SO_2), $-.196$ and .126. The other variable coefficients in this model have values at $k = .2$ which are reasonably close to the corresponding values in Table IV with the exception of the July temperature variable whose value is now .017. It is interesting to note that the sign of this coefficient is now positive, whereas in the linear model it is negative. This further substantiates a remark made at the end of Section 5 in connection with Figure 4; namely, the relative effect of mean July temperature is questionable. In this respect, the subset of variables chosen by the ridge elimination method appears to be a better choice than the subset resulting from the minimum C_p criterion.

The summary statistics for the model with fifteen variables including logarithmic transformed pollution potentials are $\varphi^* = .203$ and $R^2 = .797$ at $k = 0$; and $\varphi^* = .272$ and $R^2 = .587$ at $k = .2$. These values are slightly larger than the corresponding values given in Table IV where the pollution variables are not transformed. Residual plots do not reveal behavior sufficiently unusual to reject the statistical assumptions made in either of these two models. We conclude that the statistical evidence is not sufficiently strong to favor one of these models over the other, and that the choice should be based on physical rather than statistical arguments.

As noted in Section 2, Los Angeles and San Francisco have very large HC and NO_x pollution potential values in comparison to the other data points. By replacing the pollution potential variables with the natural logarithms, the dominating effect of these two data points is dampened, and the variables are more evenly distributed throughout their range. This can also be accomplished by simply deleting the two data points in question. The analysis of the linear model with 58 data points, rather than 60, still reveals a lack of stability in least squares estimates. The values of the coefficients of the three pollution variables at $k = 0$ and $k = .2$ respectively are: (i) HC, $-.485$ and $-.063$; (ii) NO_x, .691 and .145; (iii) SO_2, $-.130$ and .181. With two exceptions, other coefficients in this model have values at $k = .2$ which are quite close to the corresponding values in Table IV. The population per household coefficient is now $-.018$ and the relative humidity coefficient is .022. The corresponding summary statistics for the model and the residual plots are very similar to those based on the 60 observations. The relatively large increase in the coefficient associated with the NO_x term, both in the model with logarithmic transformed pollutants and the linear model with Los Angeles and San Francisco deleted, is not surprising since these two SMSA's had total mortality rates of 861.83 and 911.70, well below the sample mean of 940.36.

In summary, the ridge regression coefficients obtained at $k = .2$ for the fifteen explanatory variables listed in Table IV, with one notable exception, appear to be reasonable values upon which to base a quantification of the association between these variables and the total mortality rate. The exception is the NO_x pollution potential coefficient, which is very sensitive to both the logarithmic transformation and the deletion of two data points as previously indicated. This sensitivity highlights the need for model specification based on physical considerations as well as an appropriate expansion of the data base employed in these analyses. The

Los Angeles and San Francisco data provide information on a rather unusual combination of factors and should motivate further investigation rather than rejection of the two observations. In regards to the other two pollution potentials, the HC term consistently exhibits no detrimental effect on mortality, while the positive association of the SO_2 potential is substantial. The effects of precipitation, population density, the percent non-white, the percent under $3000, and relative humidity are all increasing, with the first three of these variables having relatively large effects. The other weather and socioeconomic variables investigated possess negative coefficients and so have a decreasing effect on the total mortality rate, with relatively large contributions coming from the January temperature and education terms.

7. Acknowledgments

The authors are particularly grateful to Miss D. Galarneau and Mr. H. Gugel for their computational aid, and to Mr. H. Ury, Mr. A. Hexter and the referees for their many helpful comments on an earlier version of this paper.

References

[1] Benedict, H. M. (1971). "Plant Damage by Air Pollutants: CRC-APRAC Project No. CAPA-2-68," presented at the Automotive Air Pollution Res. Symp., Chicago.

[2] Daniel, C., and Wood, F. S. (1971). *Fitting Equations to Data (Computer Analysis of Multifactor Data for Scientists and Engineers)*, John Wiley.

[3] Draper, N. R. (1963). "'Ridge Analysis' of Response Surfaces," *Technometrics 5*, 469–479.

[4] Duffy, E. A., and Carroll, R. E. (1967). *United States Metropolitan Mortality, 1959–1961*, PHS Publication No. 999-AP-39, U. S. Public Health Service, National Center for Air Pollution Control.

[5] Ganz, Alexander, (1968). Department of City and Regional Planning. Massachusetts Institute of Technology. Unpublished Data.

[6] Glasser, M., and Greenburg, L. (1971). "Air Pollution Mortality and Weather," *Archives Environmental Health 22*, 334–343.

[7] Green, L. W. (1970). "Manual for Scoring Socioeconomic Status for Research on Health Behavior," *Public Health Reports 85*, 815–827.

[8] Hexter, A. C., and Goldsmith, J. R. (1971). "Carbon Monoxide: Association of Community Air Pollution and Mortality," *Science 172*, 265–268.

[9] Hickey, R. J., Boyce, D. E., Harner, E. B., and Clelland, R. C. (1970). "Ecological Statistical Studies Concerning Environmental Pollution and Chronic Disease," *IEEE Transactions on Geoscience Electronics GE-8*, 186–202.

[10] Hoerl, A. E. (1962). "Application of Ridge Analysis to Regression Problems," *Chemical Engineering Progress 58*, 54–59.

[11] Hoerl, A. E. and Kennard, R. W. (1970). "Ridge Regression: Biased Estimation for Nonorthogonal Problems," *Technometrics 12*, 55–67.

[12] Hoerl, A. E. and Kennard, R. W. (1970). "Ridge Regression: Applications to Nonorthogonal Problems," *Technometrics 12*, 69–82.

[13] Hocking, R. R. and Leslie, R. N. (1967). "Selection of the Best Subset in Regression Analysis," *Technometrics 9*, 531–540.

[14] Holland, W. W., Spicer, C. C., and Wilson, J. M. G. (1961). "Influence of the Weather on Respiratory and Heart Disease," *The Lancet 2*, 338–341.

[15] LaMotte, L. R. and Hocking, R. R. (1970). "Computational Efficiency in the Selection of Regression Variables," *Technometrics 12*, 83–93.

[16] Lave, L. B. and Seskin, E. P. (1970). "Air Pollution and Human Health," *Science 169*, 723–733.

[17] Mallows, C. L. (1964). "Choosing Variables in a Linear Regression: A Graphical Aid," presented at the Central Regional Meeting of the Institute of Mathematical Statistics, Manhattan, Kansas.

[18] Marquardt, D. W. (1970). "Generalized Inverses, Ridge Regression Biased Linear Estimation, and Nonlinear Estimation," *Technometrics 12*, 591–612.

[19] NEWHOUSE, J. P. and OMAN, S. D. (1971). "An Evaluation of Ridge Estimators," Report No. R-716-PR, Rand Corp., Santa Monica, Calif.

[20] OECHSLI, F. W. and BUECHLEY, R. W. (1970). "Excess Mortality Associated with Three Los Angeles September Hot Spells," *Environmental Research 3*, 277–284.

[21] REINKE, W. A., (1969). "Multivariate and Dynamic Air Pollution Models," *Archives Environmental Health 18*, 481–484.

[22] SHY, C. M., CREASON, J. P., PEARLMAN, M. D., MCCLAIN, K. E., BENSON, F. B., and YOUNG, M. M. (1970). "The Chattanooga School Children Study: Effects of Community Exposure, to Nitrogen Dioxide. I. Methods, Description of Pollutant Exposure, and Results of Ventilatory Function Testing," *J. Air Pollution Control Assoc. 20*, 539–545.

[23] U. S. Department of Commerce (1970). Bureau of Census. *Area Measurement Reports*. Series GE-20 and Records.

[24] U. S. Department of Commerce (1966). Bureau of Census. *U. S. Census of Housing: State and Small Areas*.

[25] U. S. Department of Commerce (1960). Bureau of Census. *U. S. Census of Population: U. S. Summary*. Part A (Number of Inhabitants). Table 22.

[26] U. S. Department of Commerce (1960). Bureau of Census. *U. S. Census of Population: U. S. Summary*. Part B (General Population Characteristics). Table 63.

[27] U. S. Department of Commerce (1960). Bureau of Census. *U. S. Census of Population: U. S. Summary*. Part B (General Population Characteristics). Table 151.

[28] U. S. Department of Commerce (1960). Bureau of Census. *U. S. Census of Population: U. S. Summary*. Part B (General Population Characteristics). Table 152.

[29] U. S. Department of Commerce (1968). *Environmental Data Service. Climatic Atlas of the United States*, U. S. Govt. Printing Office.

[30] U. S. Department of Commerce (1962). Weather Bureau. *Decennial Census of United States Climate*—Monthly Normals of Temperature, Precipitation ,and Heating Degree Days (1931–1960). *Climatological Data*. Monthly and Annual; *Local Climatological Data*. Monthly and Annual.