FLIGHT TRANSPORTATION LABORATORY
DEPARTMENT OF AERONAUTICS AND ASTRONAUTICS
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

FTL Report R77-3
August 1977

AN APPLICATION OF ADVANCED STATISTICAL TECHNIQUES TO FORECAST THE DEMAND FOR AIR TRANSPORTATION

DENNIS F. X. MATHAISEL
NAWAL K. TANEJA

TABLE OF CONTENTS

I. Introduction and Objectives

For some time now regression models, often calibrated using the ordinary least-squares (OLS) estimation procedure, have become common tools for forecasting the demand for air transportation. However, in recent years more and more decision makers have begun to use these models not only to forecast traffic, but also for analyzing alternative policies and strategies. Despite this increase in scope for the use of these models for policy analysis, few analysts have investigated in depth the validity and precision of these models with respect to their expanded use. In order to use these models properly and effectively it is essential not only to understand the underlying assumptions and their implications which lead to the estimation procedure, but also to subject these assumptions to rigorous scrutiny. For example, one of the assumptions that is built into the ordinary least-squares estimation technique is that the explanatory variables should not be correlated among themselves. If the variables are fairly collinear, then the sample variance of the coefficient estimators increases significantly, which results in inaccurate estimation of the coefficients and uncertain specification of the model with respect to inclusion of those explanatory variables. As a corrective procedure, it is a common practice among demand analysts to drop those explanatory variables out of the model for which the t-statistic is insignificant. This is not a valid procedure since if collinearity is present the increase in variance of the coefficients will result in lower values of the t-statistic and rejection from the demand model of those explanatory variables which in theory do explain the variation in the dependent variable. Thus, if one or more of the assumptions underlying the OLS estimation procedure are violated, the analyst must either use appropriate correction procedures or use alternative estimation techniques.

The purpose of the study herein is three-fold: (1) develop a "good" simple regression model to forecast as well as analyze the demand for air transportation; (2) using this model, demonstrate the application of various statistical tests to evaluate the validity of each of the major assumptions underlying the OLS estimation procedure with respect to its expanded use of policy analysis; and, (3) demonstrate the application of some advanced and relatively new statistical estimation procedures which are not only appropriate but essential in eliminating the common problems encountered in regression models when some of the underlying assumptions in the OLS procedure are violated.

The incentive for the first objective, to develop a relatively simple single equation regression model to forecast as well as analyze the demand for air transportation (as measured by revenue passenger miles in U.S. Domestic trunk operations), stemmed from a recently published study by the U.S. Civil Aeronautics Board [CAB, 1976]. In the CAB study a five explanatory variable regression equation was formulated which had two undesirable features. The first was the inclusion of time as an explanatory variable. The use of time is undesirable since, from a policy analysis point of view, the analyst has no "control" over this variable, and it is usually only included to act as a proxy for other, perhaps significant, variables inadvertently omitted from the equation. The second undesirable feature of the CAB model is the "delta log" form of the equation (the first difference in the logs of the variables),which allowed a forecasting interval of only one year into the future. This form was the result of the application of a standard correction procedure for collinearity among some of the explanatory variables.

In view of these two undesirable features, it was decided to attempt to improve on the CAB model. In addition to the explanatory variables considered in the CAB study a number of other variables were analyzed to determine their appropriateness in the model. Sections II and III of this report describe the total set of variables investigated as well as a method for searching for the "best" subset. Then, Section IV outlines the decisions involved in selecting the appropriate form of the equation.

The second objective of this study is to describe a battery of statistical tests, some common and some not so common, which evaluate the validity of each of the major assumptions underlying the OLS estimation procedure with respect to single equation regression models. The major assumptions assessed in Section V of this report are homoscedasticity, normality, autocorrelation, and multicollinearity. The intent here is not to present all of the statistical tests that are available, for to do so would be the purpose of regression textbooks, but to scrutinize these four major assumptions enough to remind the analyst that it is essential to investigate in depth the validity and precision of the model with respect to its expanded use of policy analysis. It is hopeful that the procedure outlined in this report sets an example to demand modeling analysts of the essential elements used in the development of reliable forecasting tools.

The third and ultimate objective of this work is to demonstrate the use of some advanced corrective procedures in the event that any of the four above mentioned assumptions have been violated. For example, the problem of autocorrelation can be resolved by the use of generalized least-squares(GLS), which is demonstrated in Section VI of this report; and the problem of multicollinearity , usually corrected by employing the cumbersome and restrictive delta log form of equation, has been eliminated by using Ridge regression (detailed in Section VII). Finally, in Section VIII an attempt is made to determine the "robustness" of a model by first performing an examination of the residuals using such techniques as the "hat matrix", and second by the application of the recently developed estimation procedures of Robust regression. Although the techniques of Ridge and Robust regression are still in the experimental stages, sufficient research has been performed to warrant their application to significantly improve the currently operational regression models.

II. A Search for the "Best" Set of Explanatory Variables

One criterion for variable selection was to find that minimum number of independent variables which maximize the prediction and control of the dependent variable. The problem is not one of finding a set of independent variables which provides the most control for policy analysis, or those variables which best predict the behavior of the dependent variable, but, rather, constraining the number of these variables to a minimum. The hazards of using too many explanatory variables are widely known: an inevitable increase in the variance of the predicted response [Allen,1974]; a more difficult model to analyze and understand; a greater presence of highly intercorrelated explanatory variables; a reduction in the number of degrees of freedom; and a more expensive model to maintain since the value of each explanatory variable must also be predicted.

Initially,ten explanatory variables were selected for analysis based partly on the study by the Civil Aeronautics Board [CAB, 1976]. These variables are:

(1) TIME - included to provide for any trend resulting from variables omitted from the model.

(2) YLD = Average scheduled passenger yield - total domestic passenger revenues plus excess baggage charges and applied taxes divided by scheduled revenue passenger miles. The yield is deflated by the consumer price index to arrive at the real cost per mile to the consumer.

(3) GNP = Gross National Product, expressed in real terms (1967 dollars) to measure changes in purchasing power.

(4) DJLAG = Dow Jones Industrial Average , lagged by one year, computed from the quarterly highs and lows of the Dow "30" price range.

(5) UNEMPLY = Unemployment - percent of the civilian labor force.

(6) PERATIO = Price to Earnings Ratio - ratio of price index for the last day of the quarter to quarterly earnings (seasonally adjusted annual rate). Annual ratios are averages of quarterly data.

(7) GOVBND or INDBND = Bond Rate - both U.S. Government bond yields and industrial bond yields were studied.

(8) DPI = Disposable Personal Income.

(9) IIP = Index of Industrial Production.

(10) SINDEX = Quality of Service - an index obtained by principal component analysis of five variables: total available seat miles, average overall flight stage length, average on-line passenger trip length, average available seats per aircraft, and average overall airborne speed. The index is used as a measure of the improvement in the quality of service offered. (See Section III).

The data values and sources for these ten variables are given in Table 1 (Appendix A).

The problem then is how to shorten this list of independent variables so as to obtain the "best" selection - "best" in the sense of maximizing prediction and control based on a model which is appropriate not only on theoretical grounds but which can also be calibrated relatively easily. As a start, the variables should be screened before-hand in some systematic way. They should make theoretical sense in the regression equation, and they should meet the more important characteristics such as (1) measurability, (2) consistency, (3) reliability of the source of the data, (4) accuracy in the reporting of the data, (5) forecastability, and (6) controlability - that is, at least some of the variables should be under your control.

Then, to aid the investigator in his selection, automatic statistical search procedures are available (see Hocking [1972]). One such search procedure calls for an examination of all possible regression equations[1] involving the candidate explanatory variables and selection of the "best" equation according to some simple criterion of goodness of fit. In this study a statistical measure of "total squared error" was utilized as a criterion for selecting independent variables.[2] This statistic, given by Mallows [1964],has a bias component and a random error component.

The total squared error (bias plus random) for n data points, using a fitted equation with p terms (including b_0) is given by:

$$\Gamma_p = \left[\frac{1}{\sigma^2} \underbrace{\sum_{i=1}^{n} (\nu_i - \eta_i)^2}_{\text{bias component}} + \underbrace{\sum_{i=1}^{n} \sigma^2(\hat{Y}_i)}_{\text{random error component}} \right] \tag{II.1}$$

where: ν_i = E $[Y_i]$ according to the true relation

η_i = E $[Y_i]$ according to the fitted equation

$\sigma^2(\hat{Y}_i)$ = variance of the fitted value Y_i

σ^2 = true error variance.

[1] Note that the all possible regression equations search procedure presupposes that the functional form of the regression relation has already been established. Both the additive form and the multiplicative form were evaluated simultaneously. The multiplicative form (or log model) was our final choice, so it will be represented here. See Section IV.

[2] Other statistical selection criteria include step-wise regression, R_p^2, MSE_p , and t-directed search. See Neter and Wasserman [1974].

With a good unbiased estimate of σ^2(call it $\hat{\sigma}^2$) it can be shown [Daniel and Wood, 1971 ; Gorman and Tomau, 1966] that the statistic C_p is an estimator of Γ_p:

$$C_p = \frac{SSE_p}{\hat{\sigma}^2} - (n-2p) \qquad (II.2)$$

where SSE stands for the sum of squares due to error.

When there is negligible bias in a regression equation (i.e. sum of squares due to bias, SSB, is zero) the C_p statistic has an expected value of p:

$$E[C_p | SSB_p = 0] = p \qquad (II.3)$$

Thus, when the C_p values for all possible regressions are plotted against p, those equations with negligible bias will tend to cluster about the line $C_p = p$. Those equations with substantial bias will tend to fall above the line. Since bias is due to variables omitted from the equation, adding terms may reduce the bias but at the added expense of an increase in the variance of the predicted response. The trade-off is accepting some bias in order to get a lower average error of prediction. The dual objective in this study, maximum prediction with a minimum number of variables, now becomes clear. Therefore, the object was to find that set of explanatory variables that leads to the smallest C_p value. If that set should contain substantial bias another set may be preferred with a slightly larger C_p which does not contain substantial bias.

The C_p values for all possible regressions were calculated on the UCLA BMD09R regression series computer program, and are plotted against p in Figure A. The C_p value for point A is the minimum one, and appears to contain negligible bias. Equations represented by points B and C and surrounding points are also potential candidates, although point B does contain some small bias, and at point C the number of variables has increased by one.

From a purely statistical point of view, the equation represented by point A in Figure A (containing the five explanatory variables TIME, YLD, DJLAG, IIP, SINDEX) appears to represent the "best" set of independent variables. But, the solution is unique to the C_p statistical criterion, and therefore may be "best" only in terms.of this criterion. Not all of the statistical procedures for variable search (R^2 criterion, MSE criterion, foreward/backward stepwise regression, or t-directed search) may lead to the same conclusion. Indeed each procedure has its unique solution, and the conclusions are only valid for our sample data. Furthermore, the importance of subjective evaluation can not be overlooked.

Consider the model chosen on the basis of minimum C_p criterion:

$$\ln RPM = b_0 + b_1 \ln TIME + b_2 \ln YLD + b_3 \ln DJLAG + b_4 \ln IIP + b_5' \ln SINDEX \tag{II.4}$$

There is no control over the variable TIME. In fact, as is often typical in time series data, this variable is merely a surrogate for those variables omitted from the regression equation and some which are included in the equation. There is also no control over the Dow Jones index, DJLAG, and there is some question as to whether this variable can be forecasted re-

Page 10 is a blank page.

liably. The index of industrial production variable, IIP, is supposed to measure changes in the economy and relate them to the growth in air travel, but it is only a measure of the industrial or business economy. It does not reflect changes in personal spending. Since the attempt in this study is to develop an aggregate model of air travel demand, it would be more appropriate to employ an aggregate measure of the economy - for example, GNP.

What remains after this pragmatic investigation is the following selected relationship:

$$\ln \text{RPM} = b_0 + b_1 \ln \text{YLD} + b_2 \ln \text{GNP} + b_3 \ln \text{SINDEX} \qquad \text{(II.5)}$$

The number of explanatory variables has been reduced considerably and yet the analyst still remains in control over two of them - YLD and SINDEX. The first variable, YLD, tracks changes in the cost of air travel; the second variable, GNP, measures changes in passenger traffic associated with changes in the national economy; and the third variable, SINDEX, is an indicator of the service offered by the industry. Thus, the objectives in the variable selection process appear to be met:

- the number of variables have been kept to a bare minimum;
- the variables make theoretical sense in the regression equation;
- the variables are measurable, consistent, reliable, accurate, forecastable, and, with the exception of GNP, controlable, that is, under the control of the policy decision maker.

III. Development of a Quality of Service Index Through Principal Component Analysis

An important measure which should be included in the economic modeling of the growth of air transportation is the quality of service offered. The level of service tends to propagate the level of air travel: that is, as the quality of air service increases, demand increases. Suppose one was to represent the improvement in the quality of air service through the following variables.

- average available seats per aircraft (SEATS)
- average overall airborne speed (SPEED)
- average on-line passenger trip length (PTL)
- average overall flight stage length (FSL)
- total available seat miles (ASM)

If the classical supply and demand equation economic modeling principle was to be employed it would be possible to design a demand equation expressing RPM as a function of the quality of service,

$$\text{RPM} = f_1(\text{Quality of Service},\ldots),$$

and a supply equation expressing the quality of service as a function of the above five variables,

$$\text{Quality of Service} = f_2(\text{SEATS},\text{SPEED},\text{PTL},\text{FSL},\text{ASM}).$$

In this way adding an additional equation to form a simultaneous equation model allows for a complex interaction within the system conjoining the demand and supply variables. This interaction, however, is undesirable for the following reasons. First, by nature of the simultaneous equation specification, the quality of service variable will be in a dynamic state. The demand and quality of service will be allowed to interact with each other in two-way causality. Although its cause and effect role is different in the two equations, the quality of service variable is still considered to be endogenous due to its dependent role. In the present model specification,for simplicity,it was decided to represent the improvement in quality of service in a static state with causality in only one direction. Second, in a simultaneous equation model it is necessary to "identify" the quality of service function (in terms of total travel time, for example). Yet, since this is an aggregate model of air transportation demand without regard to length of haul or nature of travel (business, pleasure, personal, etc.), it would be incorrect to specify the quality of service offered in such terms as total travel time or total frequency available. Third, a simultaneous equation model is more complex than a single equation specification, which makes such a model more difficult to control and interpret - incongruous to the objective of developing a simple model. Fourth, a supply equation cannot be designed for level of service from the five variables listed since at least one of the variables, total available seat miles, is itself a supply variable. Thus, while on theoretical grounds it may be argued that the dual-equation specification is more appropriate, the intention is to be crude because the focus here is on the estimation procedures rather than the theoretical specification.

Therefore, the objective in this study of an aggregate travel demand model is not to construct a supply equation for use in a simultaneous equation model, but rather to design one factor which can be used as a proxy for the quality of service in a single equation model. Principal component analysis (PCA) [Bolch & Huang, 1974; Davis, 1973] takes the five variables mentioned above, and makes a linear combination of them in such a manner that it captures as much of the total variation as possible. This linear combination, or principal component, will serve as a proxy for level of service.

Before discussing the development of the principal component analysis technique, it will be difficult to understand the meaning of the linear combination of the five variables unless they are all measured in the same units (the raw data is given in Table 2,Appendix A). A linear combination of seats, miles and miles-per-hour is at best unclear. Therefore, before computing the components it is possible to convert the variables to standard scores, or z-scores,

$$z = \frac{X - \bar{X}}{s}$$

(where: $\bar{X}$ is the sample mean, and s is the sample standard deviation), which are pure numbers, and the problem of forming linear combinations of different measurements vanishes.

From the standard scores of m variables it is possible to compute an m x m matrix of variances and covariances. (The variance-covariance matrix of the z-scores for the five variables is simply the correlation matrix of the original data set. This is given Table 3.) From this, one can extract m eigenvectors and m eigenvalues. Recall that elements in an m x m matrix can be regarded as defining points lying on an m-dimensional ellipsoid. The eigenvectors of the matrix yield the principal axes of the ellipsoid, and the eigenvalues represent the lengths of these axes. Because a variance -covariance matrix is always symmetrical, these m eigenvectors will be orthogonal, or oriented at right angles to each other. Figure B illustrates this principle graphically for a two-dimensional matrix.

Table 3

Variance-Covariance Matrix of the z-scores of the Five Quality of Service Variables

	ASM	FSL	PTL	SEATS	SPEED
ASM	1.0	0.975	0.923	0.966	0.934
FSL	0.975	1.0	0.909	0.933	0.954
PTL	0.923	0.909	1.0	0.975	0.878
SEATS	0.966	0.933	0.975	1.0	0.913
SPEED	0.934	0.954	0.878	0.913	1.0

Principal Axes of a 2-Dimensional Ellipsoid

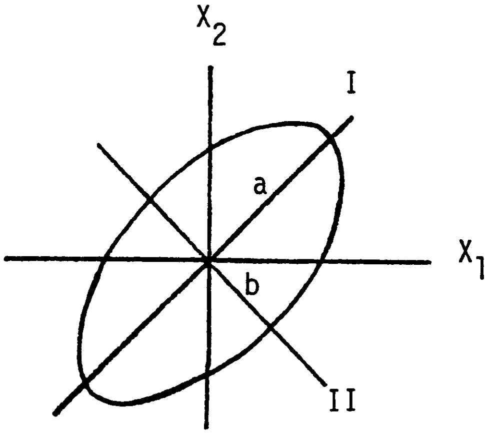


Figure B

Vectors I and II are the eigenvectors, and segments a and b are the corresponding eigenvalues. This is the case where X_1 and X_2 are positively correlated. As the correlation between the two variables increases, the ellipse more closely approaches a straight line. In the extreme case if X_1 and X_2 are perfectly positively correlated, the ellipse will degenerate into a straight line. It is not possible to demonstrate the 5-dimensional variance-covariance matrix graphically, but it can be envisioned in the same manner. The eigenvectors and eigenvalues of this 5-dimensional variance-covariance matrix can be computed easily and are shown below.

$$I = \begin{bmatrix} 0.985 \\ 0.980 \\ 0.962 \\ 0.983 \\ 0.960 \end{bmatrix} \qquad II = \begin{bmatrix} 0.041 \\ 0.139 \\ -0.251 \\ -0.151 \\ 0.222 \end{bmatrix}$$

I Eigenvalue = 4.7447 II Eigenvalue = 0.1563

$$III = \begin{bmatrix} -0.147 \\ -0.085 \\ 0.073 \\ 0.001 \\ 0.164 \end{bmatrix} \qquad IV = \begin{bmatrix} -0.062 \\ 0.111 \\ 0.079 \\ -0.093 \\ -0.033 \end{bmatrix} \qquad V = \begin{bmatrix} 0.042 \\ -0.031 \\ 0.026 \\ -0.046 \\ 0.010 \end{bmatrix}$$

III Eigenvalue = 0.0613 IV Eigenvalue = 0.0321 V Eigenvalue = 0.0057

The total of the five eigenvalues is

$$4.7447 + 0.1563 + 0.0613 + 0.0321 + 0.0057 = 5.0$$

which is equal to p, or the number of variables. This is also equivalent to the trace of the variance-covariance matrix for the z-scores. Thus, the sum of the eigenvalues represents the total variance of the data set. Note that the first eigenvalue accounts for 4.7447/5.0 = 94.89 percent of the trace. Since the eigenvalues represent the lengths of the principal axes of an ellipsoid, then the first principal axis, or <u>principal component</u>, accounts for 94.89 percent of the total variance. In other words, if one measures the variation in the data set by the first principal component, it accounts for almost 95 percent of the total variation in the observations. Principal components, therefore, are nothing more than the eigenvectors of a variance-covariance matrix.

Principal components are also a linear combination of the variables, X, of the data set. Thus

$$Y_1 = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5$$

is the first principal component, and the β's are the elements of the first eigenvector. If one were to use the standard scores in lieu of actual data, the first principal component becomes:

$$Y_1 = \overset{*}{\beta_1} z_1 + \overset{*}{\beta_2} z_2 + \overset{*}{\beta_3} z_3 + \overset{*}{\beta_4} z_4 + \overset{*}{\beta_5} z_5 .$$

where the coefficients β^* are the elements of the eigenvectors corresponding to the standardized variance-covariance matrix. Transforming this equation in terms of the original data, $Z = \frac{X - \bar{X}}{s}$, and substituting in the elements of the standardized eigenvectors, β^*, for the first principal component, the level of service index can be computed as follows:

$$\begin{aligned}\text{Level of Service Index} = {} & 0.985\left(\frac{\text{ASM} - 100.103}{88.425}\right) \\ & +0.980\left(\frac{\text{FSL} - 282.417}{95.709}\right) \\ & +0.962\left(\frac{\text{PTL} - 585.4}{81.237}\right) \\ & +0.983\left(\frac{\text{SEATS} - 77.15}{33.375}\right) \\ & +0.960\left(\frac{\text{SPEED} - 290.433}{84.987}\right)\end{aligned}$$

Therefore, this procedure results in a new set of data, termed the principal component (or factor) scores, which is a linear combination of the five level of service variables and accounts for 95 percent of the variation in the observations. The centered principal component scores are given in Table 2. To these scores the quantity 2.0 is subsequently added, which is the minimum positive integer required to transform the scores to all positive values for use in a logarithmic model. The resulting specified index, termed SINDEX, is the proxy for the quality of service offered.

IV. Functional Form of the Model

Two functional forms were investigated: the additive form,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n + \varepsilon$$

and the multiplicative form,

$$Y = \beta_0 X_1^{\beta_1} \cdot X_2^{\beta_2} \dots X_n^{\beta_n} \cdot \varepsilon.$$

The multiplicative form is in a class of models called intrinsically linear models because it can be made linear through a logarithmic transformation of the variables:

$$\ln Y = \beta_0^* + \beta_1 \ln X_1 + \beta_2 \ln X_2 + \dots + \beta_n \ln X_n + \varepsilon^*$$

where, β_0^* and ε^* are the logarithmic transformations of the constant and the error term.

The delta log model,

$$\ln Y_t - \ln Y_{t-1} = \beta_0^* + \beta_1(\ln X_{1t} - \ln X_{1t-1}) + \dots ,$$

which is commonly used because it tends to eliminate the multicollinearity and serial correlation problems,was not employed because it removes the secular trend in the data and only allows one to forecast out one period - not really suitable for long-term forecasting.

Both the additive form and the multiplicative form (or logarithmic model) were evaluated while the variable selection process was taking place (see Section II). This was necessary since the automatic variable selection technique that was used, Mallows C_p criterion, required that the functional form already be established. The log model was the final choice for the functional form. This conclusion was reached after an exhaustive and extensive evaluation of both forms during the variable selection process. To illustrate this conclusion compare the statistics of both forms using the three explanatory variables which were finally selected.

Additive Model

$$\text{RPM} = -56.556 + \underset{(0.281)}{0.934\ \text{YLD}} + \underset{(1.252)}{0.068\ \text{GNP}} + \underset{(2.914)}{28.130\ \text{SINDEX}}$$

$R^2 = 0.9780$
Standard Error of Estimate = 6.1536
n = 27

Natural Logarithm Model

$$\ln \text{RPM} = 7.433 - \underset{(-3.443)}{0.735 \ln \text{YLD}} + \underset{(1.934)}{0.592 \ln \text{GNP}} + \underset{(7.810)}{1.132 \ln \text{SINDEX}}$$

$R^2 = 0.9968$
Standard Error of Estimate = 0.0534
n = 27

The figures in paratheses are the corresponding t-ratios. Not only is the sign of the coefficient for yield in the additive model "wrong" (the plus sign of the coefficient infers that as the cost of travel increases the amount of travel increases!), but it is difficult to place any confidence in either its coefficient or that of GNP since the t-ratios are not significant.

V. Statistical Evaluation of the Ordinary Least Squares Model

In the development of an econometric air travel demand model, such as $Y = X\beta + \varepsilon$, the analyst attempts to obtain the best values of the parameters β given by estimators b. The method of ordinary least-squares produces the best estimates in terms of minimizing the sum of the squared residuals. The term "best" is used in reference to certain desired properties of the estimators. The goal of the analyst is to select an estimation procedure which produces estimators (for example price elasticity of demand) which are unbiased, efficient, and consistent [Wonnacott & Wonnacott, 1970]. The method of least-squares estimation is used in regression analysis because these properties can be achieved if certain assumptions can be made with respect to the disturbance term ε and the other explanatory variables in the model. These assumptions are:

1. Homoscedasticity - the variance of the disturbance term is constant;
2. Normality - the disturbance term is normally distributed;
3. No autocorrelation - the covariance of the disturbance term is zero;
4. No multicollinearity - matrix X is of full column rank.,

This section goes through each of these assumptions performing various tests to determine if any assumption is violated. (For further reference on each assumption see Taneja [1976]). Corrective procedures for the violations are explained and examined in subsequent sections of this report. The ordinary least-squares regression equation which is to be tested is as follows:

$$\ln RPM = 7.433 - 0.735 \ln YLD + 0.592 \ln GNP + 1.132 \ln SINDEX \qquad (V.1)$$

$$(3.507) \qquad (-3.443) \qquad (1.934) \qquad (7.810)$$

$R^2 = 0.9968$
Standard Error of Estimate = 0.0534
Durbin-Watson Statistic = 0.77
n = 27 observations
The figures in parentheses are the corresponding t - ratios.

V.1 Test for Homoscedasticity

The assumption of constant variance of the disturbance term is called homoscedasticity. When it is not satisfied, the condition is called heteroscedasticity. Heteroscedasticity is fairly common in the analysis of cross-sectional data. Air transportation demand models incorporating income distribution of air travelers, for example, usually encounter this problem because the travel behavior of the rich and poor can differ in the large and small variances of their spending pattern.

If all assumptions except the one related to homoscedasticity are valid, the estimators b produced by least-squares are still unbiased and consistent, but they are no longer minimum-variance estimators. The presence of heteroscedasticity will distort the measure of unexplained variation, thus distorting the conclusions derived from the multiple correlation coefficient R and the t - statistic tests. The problem is that the true value of a heteroscedastic residual term is dependent upon the related explanatory variables. Thus, the R and t-values become dependent on the range of the related explanatory variables. In general, the standard goodness-of-fit statistics will usually be misleading. The t and F distributed test statistics used for testing the significance of the coefficients will be overstated, and the analyst will unwarily be accepting the hypothesis of significance more often than he should.

V.1.1. Goldfeld-Quandt Test

A reasonable parametric test for the presence of heteroscedasticity is one proposed by Goldfeld and Quandt [1965]. The basic equation for n observations, $Y = X\beta + \varepsilon$, is partitioned in the form:

$$\begin{bmatrix} y_A \\ y_B \end{bmatrix} = \begin{bmatrix} x_A \\ x_B \end{bmatrix} \beta + \begin{bmatrix} \varepsilon_A \\ \varepsilon_B \end{bmatrix} \qquad \text{(V.2)}$$

The vectors and matrices with subscript A refer to the first n/2 observations, and those with subscript B refer to the last n/2 observations. If e_A and e_B represent the least-squares residuals computed for both sets of observations independently (from separate regressions),then the ratio $\frac{e'_A e_A}{e'_B e_B}$ of the sum of squares of these residuals will be distributed as $F(n/2 - p, n/2 - p)$. The null hypothesis is that the variance of the disturbance term is constant, $\sigma_A = \sigma_B$; that is, the data is homoscedastic. The null hypothesis will be rejected if the calculated F value exceeds $F(\alpha\ ; n_A - p, n_B - p)$.

The calculated F value for the model shown in equation (V.1) is

$$F_{n_A - p,\ n_B - p} = F_{13-4,\ 14-4} = \frac{SSE_A/(13-4)}{SSE_B/(14-4)}$$

$$= \frac{0.012/9}{0.008/10}$$

$$= 1.667 \qquad \text{(V.3)}$$

which is less than F(.05,9,10) = 3.02. So, the null hypothesis that the data set is homoscedastic can be accepted .

Sometimes the variance of the disturbance term changes only very slightly and heteroscedasticity is not so obvious. To help exaggerate the differences in the variance Goldfeld and Quandt suggest the omission of about 20 percent of the central observations. The elimination of 20 percent of the data reduces the degrees of freedom, but it tends to make the difference between $e'_A e_A$ and $e'_B e_B$ larger. Of the 20 total observations in the data base, the middle 6 observations were omitted. The calculated F value is

$$F_{n_A - p,\ n_B - p} = F_{10-4,11-4} = \frac{SSE_A/(10-4)}{SSE_B/(11-4)}$$

$$= \frac{.005/6}{.005/7}$$

$$= 1.167 \qquad (V.4)$$

which is less than F(.05,6,7) = 3.79. Again the null hypothesis that the data set is homoscedastic can be accepted.

V.1.2 Glejser Test

This test ,due to Glejser [1969], suggests that decisions about homoscedasticity of the residuals be taken on the basis of the statistical significance of the coefficients resulting from the regression of the absolute values of the least squares residuals on some function of X_j, where X_j is the explanatory variable which may be causing the problem. This would mean, for this example, considering fairly simple functions like:

$$\text{(i)} \quad |e| = b_0 + b_1 \ \ln \text{YLD}$$

$$\text{(ii)} \quad |e| = b_0 + b_1 \ \ln \text{GNP}$$

$$\text{(iii)} \quad |e| = b_0 + b_1 \ \ln \text{SINDEX}$$

and testing the statistical significance of the coefficients b_0 and b_1. Two relevant possibilities may then arise: (1) the case of "pure heteroscedasticity" when $b_0 = 0$ and $b_1 \neq 0$; and (2) the case of "mixed heteroscedasticity" when $b_0 \neq 0$ and $b_1 \neq 0$.

Using the above relationships the regression equations are:

$$\text{(i)} \quad |e| = \underset{(0.682)}{0.038} + \underset{(0.073)}{0.002} \ln \text{YLD}$$

$$\text{(ii)} \quad |e| = \underset{(o.641)}{0.080} - \underset{(-0.309)}{0.006} \ln \text{GNP}$$

$$\text{(iii)} \quad |e| = \underset{(5.586)}{0.044} - \underset{(-0.376)}{0.004} \ln \text{SINDEX}$$

The figures in parentheses are the corresponding t - ratios.

A test for deciding whether or not $b_0 = 0$ or $b_1 = 0$ is based on the t-statistic. The decision rule with this test statistic, when controlling the level of significance at α, is:

$$\text{if } |t_j| \leq t(1 - \alpha/2 \ ; \ n\text{-}2), \text{ then conclude } b_j = 0,$$

$$\text{if } |t_j| > t(1 - \alpha/2 \ ; \ n\text{-}2), \text{ then conclude } b_j \neq 0.$$

To draw inferences regarding both $b_0 = 0$ and $b_1 =0$ *jointly*, a test can be conducted which is based on the F-distribution (see Neter & Wasserman

[1974]). Define the calculated F-statistic as:

$$F_{calculated} = \frac{n(b_o - \beta_o)^2 + 2(\Sigma X_i)(b_o - \beta_o)(b_1 - \beta_1) + (\Sigma X_i^2)(b_1 - \beta_1)^2}{2(MSE)}$$

where $\beta_o \equiv 0, \beta_1 \equiv 0$, and MSE is the mean square error.

The decision rule with this F-statistic is:

if $F_{calculated} \leq F(1 - \alpha; n-2)$, then conclude that $b_o = 0$ and $b_1 = 0$ jointly .

At a 5 percent level of significance, where t(.975, 27-2) = 2.06 and F(.95, 2, 27-2) = 3.38, the results of the significance tests on b_o and b_1 are the following:

For regresssion (i) : $|t_o| = 0.862 < 2.06$

$|t_1| = 0.073 < 2.06$

$$F_{calculated} = \frac{27(0.038 - 0)^2 + 2(51.234)(0.038 - 0)(.002-0) + (98.118)(.002)^2}{2(.0075)}$$

$$= 3.127 < 3.38$$

which infer that $b_o = 0$, $b_1 = 0$, and both b_o and b_1 are equal to zero jointly. So neither pure nor mixed heteroscedasticity are associated with the variable ln YLD.

For regression (ii): $|t_o| = 0.641 < 2.06$

$|t_1| = 0.309 < 2.06$

$F_{calculated} = 2.536 < 3.38$

inferring that $b_0 = 0$, $b_1 = 0$, and both b_0 and b_1 are equal to zero jointly. So neither pure nor mixed heteroscedasticity are associated with ln GNP.

For regression (iii):

$$t_0 = 5.586 > 2.06$$

$$t_1 = 0.376 < 2.06$$

$$F_{\text{calculated}} = 3.86 > 3.38$$

Although b_0 is not equal to zero and the F-distributed statistic fails the test, b_1 <u>is</u> equal to zero- concluding once again that neither pure nor mixed heteroscedasticity are associated with the variable ln SINDEX.

V.1.3. Spearman Coefficient of Rank Correlation Test

Yet another test for heteroscedasticity was suggested by P.A. Gorringe in an unpublished M.A. thesis at the University of Manchester (see Johnston [1972]). This test computes the Spearman coefficient of rank correlation between the absolute values of the residuals and the explanatory variable with which heteroscedasticity might be associated.

In this test the variables $|e|$ and X are replaced by their ranks $|e|'$ and X', which range over the values 1 to n. The coefficient of rank correlation is the simple product-moment correlation coefficient computed from $|e|'$ and X':

$$r = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n^3 - n} \qquad (V.5)$$

where $d_i = |e_i|' = X_i'$

If heteroscedasticity is associated with the variable X, then as X changes, |e| moves in the same direction, the distance between |e|' and X' is negligible, and r approaches 1.0. Thus, for r approaching 1.0 the Spearman rank correlation test indicates severe heteroscedasticity with X ; and, for r approaching 0 the test indicates no heteroscedasticity with X. The calculations of r for the three variables ln YLD, ln GNP and ln SINDEX are given in Table 4.

$$r_{\ln YLD} = 0 \text{ , } r_{\ln GNP} \approx 0 \text{ and } r_{\ln SINDEX} \approx 0$$

thus, no heteroscedasticity is associated with any of our variables.

Table 4

Spearman Coefficient of Rank Correlation Test for Heteroscedasticity

n	|e|'	ln YLD'	d_i^2	ln GNP'	d_i^2	ln SINDEX'	d_i^2
1	23	27	16	1	484	1	484
2	19	26	49	2	289	2	289
3	7	25	324	3	16	3	16
4	18	24	36	4	196	4	196
5	13	23	100	6	49	5	64
6	8	19	121	5	9	6	4
7	14	17	9	7	49	7	49
8	20	16	16	8	144	8	144
9	11	14	9	10	1	9	4
10	5	15	100	9	16	10	25
11	27	18	81	11	256	11	256
12	21	20	1	12	81	12	81
13	1	22	441	13	144	13	144
14	6	21	225	14	64	14	64
15	12	13	1	15	9	15	9
16	2	12	100	16	196	16	196
17	9	11	4	17	64	17	64
18	26	10	256	18	64	18	64
19	24	9	225	19	25	19	25
20	15	7	64	20	25	20	25
21	3	6	9	22	361	21	324
22	22	4	324	21	1	22	0
23	25	8	289	23	4	23	4
24	10	5	25	24	196	24	196
25	17	2	225	27	100	26	81
26	4	3	1	26	484	25	441
27	16	1	225	25	81	27	121
			3276		3408		3370

$$r = 1 - \frac{6(3276)}{27^3 - 27} = 0 \qquad r = 1 - \frac{6(3408)}{27^3 - 27} \cong 0 \qquad r = 1 - \frac{6(3370)}{27^3 - 27} \cong 0$$

V.2 Test for Normality

For the model shown by Equation (V.1), it was not necessary to make an assumption that the disturbance term follow a particular probability distribution, normal or otherwise. Although this assumption is not necessary [Murphy, 1973], it is common and convenient for almost all of the test procedures used. For example, the Durbin-Watson test assumes that the disturbance terms are normally distributed. The normality assumption can be justified if it can be assumed that many explanatory variables, other than those specified in the model, affect the dependent variable. Then, according to the Central Limit Theorem (for a reasonably large sample size the sample mean is normally distributed), it is expected that the disturbance term will follow a normal distribution. However, for the Central Limit Theorem to apply, it is also necessary to assume that these factors which influence the dependent variable Y must be independent. Thus, while the assumption of normal distribution for the disturbance term is quite common and convenient, its validity is not so obvious.

Small departures from normality do not create any serious problems, but major departures should be of concern. Normality of error terms can be studied in a variety of ways.

V.2.1. A Simple Test

One test for normality is to determine whether about 68 percent of the standardized residuals, $e_i/\sqrt{MSE}$, fall between -1 and +1, or about 90 percent between -1.64 and +1.64. For the regression model under investigation, 17/27 or 63 percent of the standardized residuals fall between -1 and +1; and 25/27 or 93 percent fall between -1.64 and +1.64.

V.2.2. Normal Probability Paper

Another possibility is to plot the standardized residuals on normal probability paper. A characteristic of this type of graph paper is that a normal distribution plots as a straight line. Substantial departures from a straight line are grounds for suspecting that the distribution is not normal.

Figure C is a normal probability plot of our model generated by the UCLA BMD regression series computer program. The values of the residuals from the regression equation are laid out on the abscissa. The ordinate scale refers to the values which would be expected if the residuals were normally distributed. While not a perfect straight line, the plot of the residuals is reasonably close to that of a standard normal distribution to suggest that there is no strong evidence of any major departure from normality..

V.2.3. Chi-Square Goodness of Fit Test

In making this test one determines how closely the observed (sample) frequency distribution fits the normal probability distribution by comparing the observed frequencies to the expected frequencies. The observed frequencies (calculated from the standardized residuals) are given in the second column of Table 5. The expected frequencies from the normal distribution are given in the third column. The remainder of the table is used to compute the value of the chi-square. The further the expected values from the observed values (the larger the χ^2 value is), the poorer the fit of the hypothesized distribution. In the present case χ^2 is equal to 1.00475. This value is not

Figure C

Normal Probability Plot of Residuals

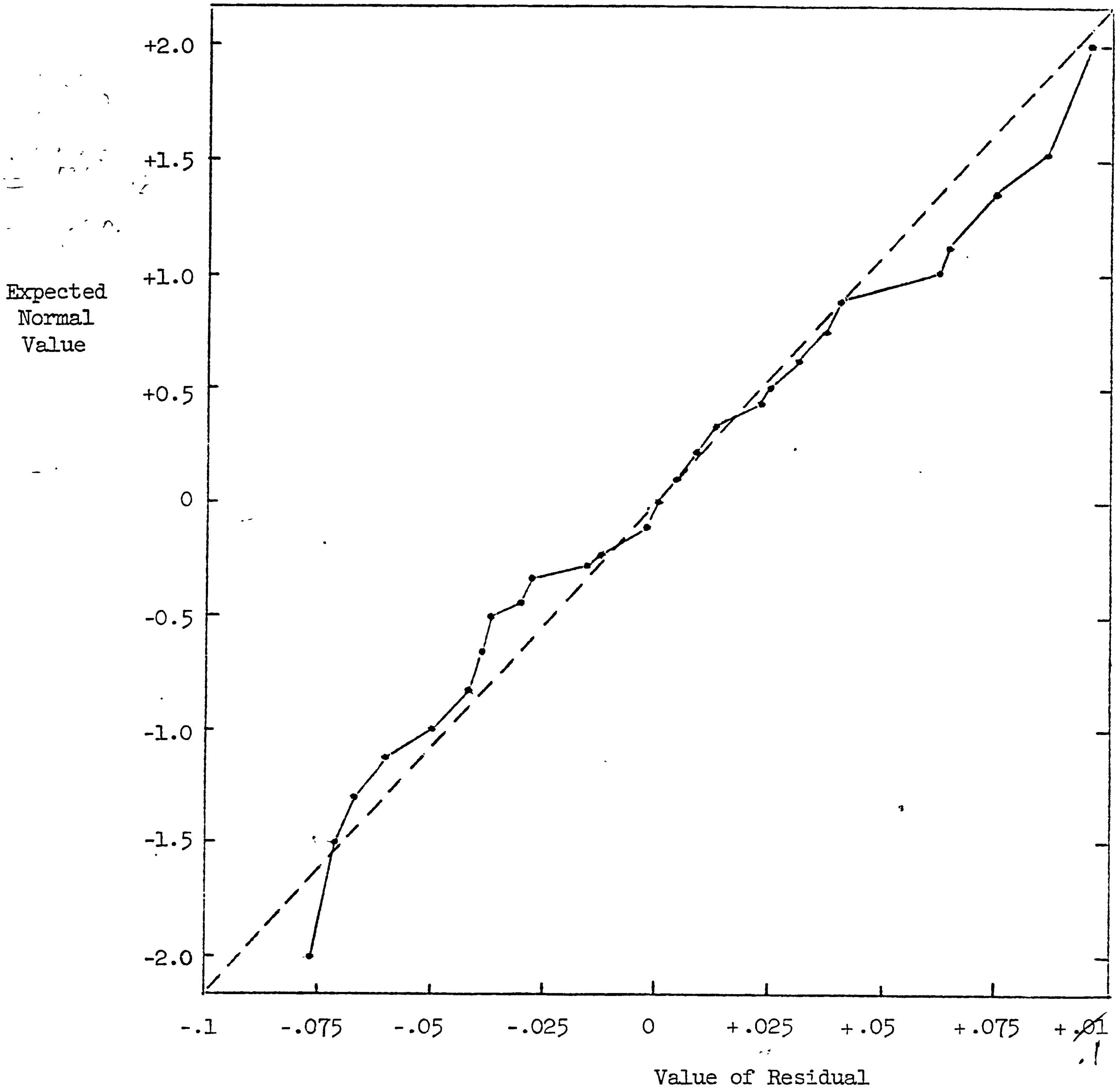


Table 5

Computation of Chi-Square

Range	Observed Frequency O_i	Expected Frequency E_i	$\frac{(O_i - E_i)^2}{E_i}$
$-\infty$ to -1.0	5	4.2849	0.11934
-1.0 to -0.5	5	4.0446	0.22568
-0.5 to 0	4	5.1705	0.26498
0 to $+0.5$	5	5.1705	0.00562
$+0.5$ to $+1.0$	3	4.0446	0.26979
$+1.0$ to $+\infty$	5	4.2849	0.11934
	27	27	$\sum_{i=1}^{6} \frac{(O_i - E_i)^2}{E_i} = 1.00475$

Table 6

Computation of the Kolmogorov-Smirnov Test Statistic D

Observed Frequency O_i	Cumulative O_i	Expected Frequency E_i	Cumulative E_i	Absolute Difference
5	.185185	4.2849	0.1587	.026485
5	.370370	4.0446	0.3085	.06187 = D
4	.5185	5.1705	0.5	.0185
5	.7037	5.1705	0.6915	.0122
3	.8148	4.0446	0.8413	.0005
5	1.0	4.2849	1.0	0

significant even at the 0.05 probability level ($\chi^2 = 11.07$), and it can be concluded that the hypothesized normal distribution fits the data satifactorily.

V.2.4 Kolmogorov-Smirnov Test

Where the chi-square test requires a reasonably large sample, the Kolmogorov-Smirnov test can be used for smaller samples. The test uses a statistic called D, which is computed as follows:

Step 1 - Compute the cumulative observed relative frequencies.

Step 2 - Compute the cumulative expected relative frequencies.

Step 3 - Compute the absolute difference of both frequencies.

Step 4 - The value of D is the maximum difference:

$$D = \max_t \left| \sum_{i=0}^{t} (O_i/n) - \sum_{i=0}^{t} (E_i/n) \right| .$$

Step 5 - Compare D to critical values from the Kolmogorov-Smirnov tables.

The larger the value of D, the less likely the hypothesized normal distribution is suitable. Tables of the statistic D exist which give the critical values for various probabilities and sample sizes [Miller, 1956]. For this case the critical value of D at a significance level of 0.10 and n=5 is 0.44698. Since the calculated value of D (0.06187) is less than the critical value, it is not significant and we again accept that our error terms are distributed normally.

V.3 Test for Autocorrelation

The third assumption in the least-squares model $Y = X\beta + \varepsilon$ represented by Equation (V.1) was that the error terms are uncorrelated random variables. This assumption of independence is equivalent to assuming zero covariance between any pair of disturbances. A violation, referred to as autocorrelation, occurs if any important underlying cause of the error has a continuing effect over several observations. It is likely to exist if the error term is influenced by some excluded variable which has a strong cyclical pattern throughout the observations. If omitted variables are serially correlated, and if their movements do not cancel out each other, there will be a systematic influence on the disturbance term. Besides the omission of variables, autocorrelation can exist because of an incorrect functional form of the model, the possibility of observation errors in the data, the estimation of missing data by averaging or extrapolating, and the lagged effect of changes distributed over a number of time periods [Huang, 1970].

The major impact of the presence of autocorrelation is its role in causing unreliable estimation of the calculated error measures. Goodness-of-fit statistics, such as the coefficient of multiple determination R^2, will have more significant values than may be warranted. In addition, the sample variances of individual regression coefficients, such as price elasticity of demand, will be seriously underestimated, thereby causing the t-tests to produce wrong conclusions. Finally, the presence of significant autocorrelation will produce inefficient estimators [Elliott, 1973].

V.3.1 The Durbin-Watson Statistic

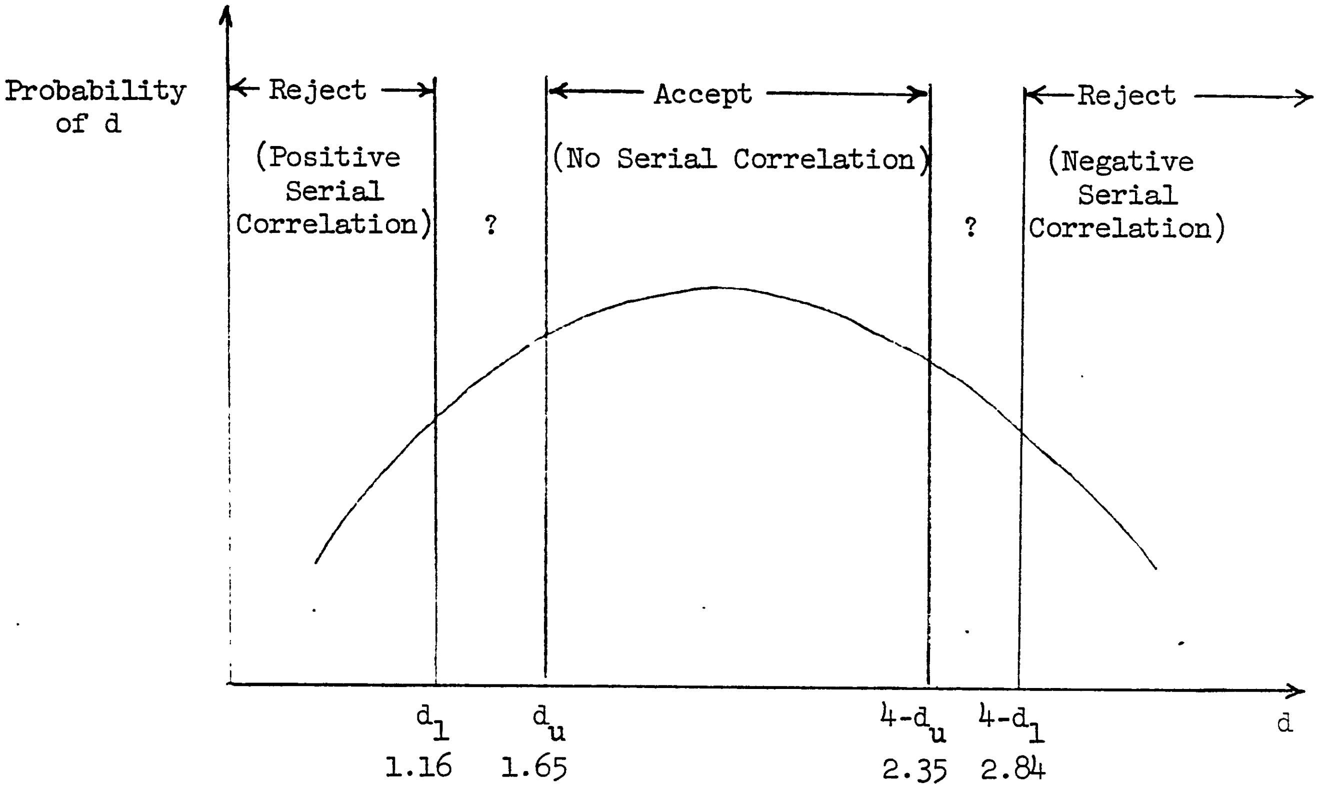
Autocorrelation of the residuals is tested by computing and interpreting the Durbin-Watson statistic [Durbin-Watson, 1950-51]. This statistic, d, is defined as follows:

$$d = \frac{\sum_{i=2}^{n} (e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}$$

where e_i is the calculated error term resultig from the estimated regression equation. A positively autocorrelated error term will result in a disproportionately small value of the d statistic since the positive and negative e_i values tend to cancel out in the computation of the squared successive differences. A negative autocorrelated error term, however, will produce large values of d as the systematic sign changes cause successive terms to add together. If no autocorrelation is present, d will have an expected value of approximately two. Two critical values of the d statistic, d_l and d_u, obtained from Durbin-Watson tables, are involved in applying the test. The critical values d_l and d_u describe the boundaries between the acceptance and rejection region for positive autocorrelation; and, the critical values $4-d_l$ and $4-d_u$ define the boundaries for negative autocorrelation. For any remaining region the test is inconclusive. The positive and negative serial correlation regions, applicable for a 5 percent level of significance and the problem presented here, are shown in Figure D [Elliott, 1973].

Figure D

Acceptance and Rejection Regions
of the Durbin-Watson Statistic



Since the Durbin-Watson statistic for the model given by Equation (V.1) is equal to 0.77 it can be concluded that the residuals are positively autocorrelated.

V.4 Test for Multicollinearity

The fourth assumption built into the econometric model of $Y = X\beta + \varepsilon$ is that matrix X is of full rank. If this condition is violated, then the determinant $|X'X|$ is equal to zero, and the ordinary least-squares estimate $b = (X'X)^{-1} X'Y$ cannot be determined. In most air travel demand models the problem is not as extreme as the case of $|X'X| = 0$. However, when the columns of X are fairly collinear such that $|X'X|$ is near zero, then the variance of b, which is equal to $\sigma^2(X'X)^{-1}$, can increase significantly. This problem is known as multicollinearity. The existence of multicollinearity results in inaccurate estimation of the parameters β because of the large sample variances of the coefficient estimators, uncertain specification of the model with respect to the inclusion of variables, and extreme difficulty in interpretation of the estimated parameters b.

The effects of multicollinearity on the specification of the air travel demand model are serious. For instance, it is a common practice among demand analysts to drop those explanatory variables out of the model for which the t-statistic is below the critical level for a given sample size and level of significance. This is not a valid procedure since if multicollinearity is present, the variances of the regression coefficients under consideration will be substantially increased, and result in lower values of the t-statistic. Thus, this procedure can result in the rejection from the demand model of those explanatory variables which in theory do explain the variation in the dependent variable Y.

V.4.1 Correlation Matrix

In air travel demand models the problem is not so much in detecting the existence of multicollinearity but in determining its severity. The seriousness of multicollinearity can usually be examined in the correlation coefficients of the explanatory variables. How high can the correlation coefficient reach before it is declared intolerable? This is a difficult question to answer, since it varies from case to case, and among different analysts. It is sometimes recommended, however, that multicollinearity can be tolerated if the correlation coefficient between any two explanatory variables is less than the square root of the coefficient of multiple determination [Huang, 1970]. An examination of the correlation matrix for the sample case (Table 7) clearly indicates significant correlation between the explanatory variables. The analyst should be cautious, however, in examining multicollinearity by using the simple correlation coefficients since such a rule is useful only for pairwise considerations; for instance, even if two columns may not be highly collinear, the determinant of (X'X) will be zero if several columns of X can be combined in a linear combination to equal another.

Table 7
Correlation Matrix

	log YLD	log GNP	log SINDEX	log RPM
log YLD	1.0			
log GNP	-0.9645	1.0		
log SINDEX	-0.9561	0.9906	1.0	
log RPM	-0.9692	0.9933	0.9966	1.0

V.4.2 Eigenvalues

The degree of multicollinearity can also be reflected in the eigenvalues of the correlation matrix. To demonstrate this it is necessary to represent the regression model in matrix notation. Let $Y = X\beta + \varepsilon$ be the regression model where it is assumed that X is (n x p) and of rank p,

β is (p x 1) and unknown, $E[\varepsilon] = 0$, and $E[\varepsilon\varepsilon'] = \sigma^2 I_n$. Let $\hat{\beta}$ be the least squares estimate of β, so that:

$$\hat{\beta} = (X'X)^{-1} X'Y \qquad (V.6)$$

Let X'X be represented in the form of a correlation matrix. If X'X is a unit matrix,

$$X'X = \begin{bmatrix} 1 & & & & \\ & 1 & & & \\ & & \cdot & & \\ & & & \cdot & \\ & & & & \cdot \; 1 \end{bmatrix} = I_n \quad ,$$

then the correlations among the X are zero, multicollinearity does not exist, and the variables are said to be orthogonal. However, if X'X is not nearly a unit matrix, multicollinearity does exist, and the least squares estimates, $\hat{\beta}$, are sensitive to errors. To demonstrate the effects of this condition Hoerl and Kennard [1970] propose looking at the distance between the estimate $\hat{\beta}$ and its true value β. If L is the distance from $\hat{\beta}$ to β, then the squared distance is:

$$L^2 = (\hat{\beta} - \beta)'(\hat{\beta} - \beta) \qquad (V.7)$$

Analogous to the fact that $E[\varepsilon\varepsilon'] = \sigma^2 I_n$, the expected value of the squared distance L^2 can be represented as:

$$E[L^2] = \sigma^2 \text{ Trace } (X'X)^{-1} \qquad (V.8)$$

If the eigenvalues of X'X are denoted by $\lambda_{MAX} = \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p = \lambda_{MIN} > 0$

then the average value of the squared distance from $\hat{\beta}$ to β is given by:

$$E[L^2] = \sigma^2 \sum_1^p (1/\lambda_i) > \sigma^2/\lambda_p \qquad (V.9)$$

This relationship indicates that as λ_p decreases, $E[L^2]$ increases. Thus, if X'X results in one or more small eigenvalues, the distance from $\hat{\beta}$ to β will tend to be large, and large absolute coefficients are one of the effects or errors of nonorthogonal data.

The X'X matrix in correlation form was given in Table 7. The resulting eigenvalues of X'X are:

$$\lambda_{YLD} = 2.933$$
$$\lambda_{GNP} = 0.059$$
$$\lambda_{SINDEX} = 0.008$$

Note the smallness of λ_{YLD} and λ_{SINDEX}, which reflects the multicollinearity problem. The sum of the reciprocals of the eigenvalues is

$$\Sigma(1/\lambda_i) = 1/2.933 + 1/0.059 + 1/0.008 = 142.29$$

Thus, equation (V.9) shows that the expected squared distance between the coefficient estimate $\hat{\beta}$ and β is 142.29 σ^2, which is much greater than what it would be for an orthogonal system (3 σ^2).

V.4.3 Farrar and Glauber Test to Identify Variables Involved in Multicollinearity

To identify which individual variables are most affected by collinearity, an F-distributed statistic proposed by Farrar and Glauber [1967], tests the null hypothesis (H_o: X_j is not affected) against the alternative

(H_1: X_j is affected).

Recalling that (X'X) is the matrix of simple correlation coefficients for X, as interdependence among the explanatory variables grows, the correlation matrix (X'X) approaches singularity, and the determinant |X'X| approaches zero. Conversely, |X'X| close to one implies a nearly orthogonal independent variable set. In the limit, perfect linear dependence among the variables causes the inverse matrix $(X'X)^{-1}$ to explode, and the diagonal elements of $(X'X)^{-1}$ become infinite. Using this principle, Farrar and Glauber define their test statistic as:

$$F_{(n-p,\ p-1)} = (r^{*j} - 1)\,\frac{(n-p)}{(p-1)} \qquad \text{(V.10)}$$

where r^{*j} denotes the j^{th} diagonal element of $(X'X)^{-1}$, the inverse matrix of simple correlation coefficients. The null hypothesis (H_0: X_j is not affected) is rejected if the calculated F-distributed statistic exceeds the critical F(α; n-p, p-1). The results for the model of Equation (V.1) are given in Table 8. The calculated F-statistics,

$$F_{\ln\ YLD} = 110.02\ ,\quad F_{\ln\ GNP} = 497.23\ ,\ \text{and}\quad F_{\ln\ SINDEX} = 402.37\ ,$$

are all much greater than the critical F(.05, 27-4, 4-1) = 8.64. Thus, the null hypothesis can be rejected and the alternate, that all variables are significantly affected by multicollinearity, can be accepted.

Table 8

Farrar and Glauber Test for Multicollinearity

Simple Correlation Matrix

$$
(X'X) = \begin{array}{ccc|c}
\ln \text{YLD} & \ln \text{GNP} & \ln \text{SINDEX} & \\
\left[\begin{matrix} 1.0 \\ -0.964 \\ -0.956 \end{matrix}\right. & \begin{matrix} -0.964 \\ 1.0 \\ 0.991 \end{matrix} & \left.\begin{matrix} -0.956 \\ 0.991 \\ 1.0 \end{matrix}\right] & \begin{matrix} \ln \text{YLD} \\ \ln \text{GNP} \\ \ln \text{SINDEX} \end{matrix}
\end{array}
$$

Determinant = 0.0013

Inverse Correlation Matrix

$$
(X'X)^{-1} = \begin{bmatrix} r^* \ln \text{YLD} & - & - \\ - & r^* \ln \text{GNP} & - \\ - & - & r^* \ln \text{SINDEX} \end{bmatrix} = \begin{bmatrix} 14.344 & 13.329 & 0.511 \\ 13.329 & 65.828 & -52.466 \\ 0.511 & -52.466 & 53.461 \end{bmatrix}
$$

$$F_{\ln \text{YLD}} = (14.344 - 1)\,\frac{(27-4)}{(4-1)} = (14.344)(7.67) = 110.02$$

$$F_{\ln \text{GNP}} = (65.828 - 1)(7.67) = 497.23$$

$$F_{\ln \text{SINDEX}} = (53.461 - 1)(7.67) = 402.37$$

V.5 Summary Evaluation of the Ordinary Least-Squares Model

The method of least-squares produces estimators which are unbiased, efficient and consistent if the assumptions regarding homoscedasticity, normality, no autocorrelation, and no multicollinearity can be made with respect to the disturbance term and the other explanatory variables in the model. This section tested each of these assumptions on the ordinary least-squares model of Equation (V.1). The conclusions are:

. the sample data is homoscedastic,

. the distribution of error terms is normal

. the residuals are positively autocorrelated,

. severe multicollinearity exists among all explanatory variables.

Thus, the least-squared model failed on the autocorrelation and multicollinearity tests.

The autocorrelation problem may adversely affect the property of the least-squares technique to produce efficient (minimum-variance) estimates. This means that the resulting computed least squares plane cannot be relied on, and the goodness-of-fit statistics, such as the R^2 and the t-test, will have more significant values than may be warranted. The existence of severe multicollinearity compounds these problems. When the columns of X are fairly collinear such that the determinant $|X'X|$ is near zero, then the variance of the estimators b, which is equal to $\sigma^2(X'X)^{-1}$ can increase significantly. This results in inaccurate estimation of parameters β because of the large sample variances of the coefficient estimators, uncertain specification of the model with respect to the inclusion of variables, and extreme difficultly in interpretation of the estimated parameters b.

In the sections that follow, the autocorrelation problem will be corrected by the method of generalized least-squares, and the problem of severe multicollinearity will be dealt with using a relatively new (1970) technique called "ridge regression".

VI. Correction for Autocorrelation - Generalized Least Squares

The problems associated withthe presence of autocorrelation (inflated R^2, unreliable estimates of the coefficients, inefficient estimators) are already well documented [Taneja, 1976].Standard procedures, such as the Cochrane-Orcutt [1949] iterative process, are available for correcting autocorrelation, but the procedure, which usually requires a first or second difference transformation of the variables:

$$Y - \hat{\rho}\,\tilde{Y} = (X - \hat{\rho}\,\tilde{X}) + (e - \hat{\rho}\,\tilde{e}) \qquad \text{(VI.1)}$$

where $\hat{\rho}$ is ~~estimated~~ from $e = \hat{\rho}\,e$,
and $\tilde{e}$ is the vector e lagged one period,

restricts the forecasting interval to one period - which is very limiting for long term forecasting. Another method, known as generalized least squares, corrects for autocorrelation and is not as restrictive as the Cochrane-Orcutt process.

The classical ordinary least squares (OLS) linear estimation procedure is characterized by a number of assumptions concerning the error term in the regression model. Specifically, the disturbance term ε in

$$Y = X\beta + \varepsilon \qquad \text{(VI.2)}$$

is supposed to satisfy the following requirement:

$$E[EE'] = \sigma^2 I_n = \sigma^2 \begin{bmatrix} 1 & & & & \\ & 1 & & & \\ & & 1 & & \\ & & & \ddots & \\ & & & & 1 \end{bmatrix} \qquad \text{(VI.3)}$$

where the identity matrix I_n is of order (n x n). Note that in the OLS model the variances of the disturbance terms are constant and the covariances are zero, $COV(\varepsilon_i \varepsilon_j) = 0$, for $i \neq j$. If the disturbance terms at different observations are dependent on each other, then the dependency is reflected in the correlation of error terms with themselves (i.e. in the covariance terms) in preceding or subsequent observations. This autocorrelation of disturbance terms violates the assumption given by Equation (VI.3).

If this assumption (VI.3) is not made, but all the other assumptions of the OLS model are retained,the result is a generalized least squares (GLS) regression model. The GLS model is:

$$Y = X\Gamma + \varepsilon , \tag{VI.4}$$

for which the variances of the disturbance terms are still constant but the covariances are no longer zero, $COV(\varepsilon_i \varepsilon_j) \neq 0$, for $i \neq j$. Consequently, the variance-covariance matrix of disturbances for the first-order autoregressive process

$$\varepsilon_t = \rho \, \varepsilon_{t-1} + \upsilon \tag{VI.5}$$

(where υ is distributed as $N(0, \sigma^2 I)$ and $|\rho| < 1$) is given by [Kmenta, 1971]:

$$E[\varepsilon\varepsilon'] = \sigma_\varepsilon^{\,2} \, \Omega \tag{VI.6}$$

$$= \sigma_\varepsilon^{\,2} \begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{t-1} \\ \rho & 1 & \rho & \cdots & \rho^{t-2} \\ \rho^2 & \rho & 1 & \cdots & \rho^{t-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{t-1} & \rho^{t-2} & \rho^{t-3} & \cdots & 1 \end{bmatrix} \tag{VI.7}$$

In OLS the estimate of β is given by:

$$\hat{\beta} = (X'X)^{-1} X'Y . \qquad (VI.8)$$

In GLS the estimate of Γ is given by:

$$\hat{\Gamma} = (X'\Omega^{-1}X)^{-1} \; (X'\Omega^{-1}Y). \qquad (VI.9)$$

The preceding discussion has been carried out on the assumption that Ω, the variance-covariance matrix of the disturbances, is known. In many cases Ω is not known. But if the disturbances follow a first-order autoregressive scheme,

$$\varepsilon_t = \rho\varepsilon_{t-1} + \upsilon \qquad (VI.10)$$

then Ω involves only two unknown parameters, σ^2 and ρ, which can be readily estimated.

A version of the generalized least squares regression technique is programmed into the TROLL computer software system developed at the National Bureau of Economic Research. TROLL automatically performs an iterative search for the best value of the ρ parameter. The results of this search are:

$$\log RPM = \underset{(1.291)}{2.133} - \underset{(-3.646)}{0.794} \log YLD + \underset{(6.121)}{1.467} \log GNP + \underset{(4.731)}{0.652} \log SINDEX$$

$R^2 = 0.9757$ (VI.11)

Standard Error of Estimate = 0.0349

n = 27

Durbin-Watson = 1.89

$\rho = 0.8796$

The Durbin-Watson statistic (1.89) is not significant even at the .05

probability level (see Section V.3.). Thus, serial correlation of the error terms has been corrected, and it was not necessary to restrict the forecasting interval to one period as in the case of using the Cochrane-Orcutt technique. Also, the t-ratios of all coefficients remain significant - in fact the t-ratios for GNP and YLD show an improvement over the OLS results.

VII. Correction for Multicollinearity - Ridge Regression

In Section V.4.2 the use of eignenvalues of the X'X matrix in correlation form was presented to diagnose the problem of multicollinearity. It was shown that the average value of the squared distance from $\hat{\beta}$, the least squares estimated coefficient, to β, the true value, is given by:

$$E[L^2] = \sigma^2 \, \mathrm{Trace}(X'X)^{-1} \qquad \text{(VII.1)}$$

$$= \sigma^2 \sum_{1}^{P} (1/\lambda_i) > \sigma^2/\lambda_p \qquad \text{(VII.2)}$$

where $\lambda_{MAX} = \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p = \lambda_{MIN} > 0$ denote the eigenvalues of X'X. This indicates that if one or more of the eigenvalues of X'X is small, the distance from $\hat{\beta}$ to β will tend to be large - that is, the coefficients will be inflated. This pulling of the least squares estimates away from the true coefficients is the effect of nonorthogonality of the prediction variables. Indeed, this was the case for the data when it was found that $E[L^2]=142.29\ \sigma^2$. If multicollinearity did not exist $E[L^2]$ would be $3\sigma^2$!

A.E. Hoerl [1962] first suggested that to control the inflation and general instability of the least squares estimates resulting from collinear data, one can use a procedure based on adding small positive quantities to the diagonal of X'X. This can be represented mathematically by adding kI, $k \geq 0$, to the matrix X'X in the least squares estimate

$$\hat{\beta} = (X'X)^{-1}X'Y \qquad \text{(VII.3)}$$

to give:

$$\hat{\beta}_k = (X'X + kI)^{-1}X'Y \qquad k \geq 0. \qquad \text{(VII.4)}$$

Equation (VII.4) has mathematical similarities to the determination of the maximum and minimum boundaries or "ridges" of quadratic response functions [Hoerl,1964]. Thus, regression analysis built around Equation (VII.4) has been labeled "ridge regression" [Hoerl & Kennard, 1970].

Generally speaking, the estimate β_k given by ridge regression provide a procedure by which one can examine the sensitivity of the coefficients to the effects of collinear data. The inflated coefficients are mathematically being required to decrease while simultaneously obtaining the best possible fit to the data. By varying the value of k in Equation (VII.4) it is possible to examine how stable the analytic solution of $\hat{\beta}_k$ is and simultaneously determine if the best solution is stable. Hoerl describes the procedure this way:

> Basically, the principle which is being applied is as follows: Given the analytic solution the ridge analysis determines the unique combinations of solutions which minimize the lack of fit of the data while decreasing the size of the coefficients. In addition, a solution is determined which not only "fits" the data, but is simultaneously a stable solution - the real crux of the matter.[3]

Since the ridge estimator contains the parameter k,which must be fixed before the value of the estimator is determined, a major problem lies in determining a reasonable value for k. Hoerl and Kennard propose a scheme for choosing an appropriate k by plotting the individual coefficients$(\hat{\beta}_{k1},\hat{\beta}_{k2},\ldots,\hat{\beta}_{kp})$ against k and noting the value of k for which the coefficients stabilize. This two-dimensional protrayal is called a"Ridge Trace" [Hoerl & Kennard, 1970].

[3] Arthur E. Hoerl, "Application of Ridge Analysis to Regression Problems", Chemical Engineering Progress, Volume 58, No. 3, March 1962, p.58.

FigureE shows the Ridge Trace for this problem. This trace was constructed by computing a total of 9 regressions and 9 values of k in the interval (0.5) indicated by the dots. The Ridge Trace was produced using the TROLL &RIDGE macro developed by the Computer Research Center of the National Bureau of Economic Research. &Ridge solved Equation (VII.4) for k=0 and for each user-specified value of k. Both X and Y were centered (using weighted means $\overline{y} = \Sigma w_i y_i/\Sigma w_i$ and $\overline{x} = \Sigma w_i x_{ij}/\Sigma w_i$), and scaled (using a weighted length for the columns of X, $S_j = \Sigma w_i(x_{ij}-\overline{x}_j)^2$, and $S_y = \Sigma w_i(y_i-\overline{y})^2$ for Y.) In addition, a prior vector δ of the coefficients, estimated from the generalized least squares correction for autocorrelation (see Section VI), was provided. The ridge estimate with a non-zero prior δ for the coefficient vector is:

$$\hat{\beta}_k = \delta + (X'X + kI)^{-1} X'(Y-X\delta) \qquad \text{(VII.5)}$$

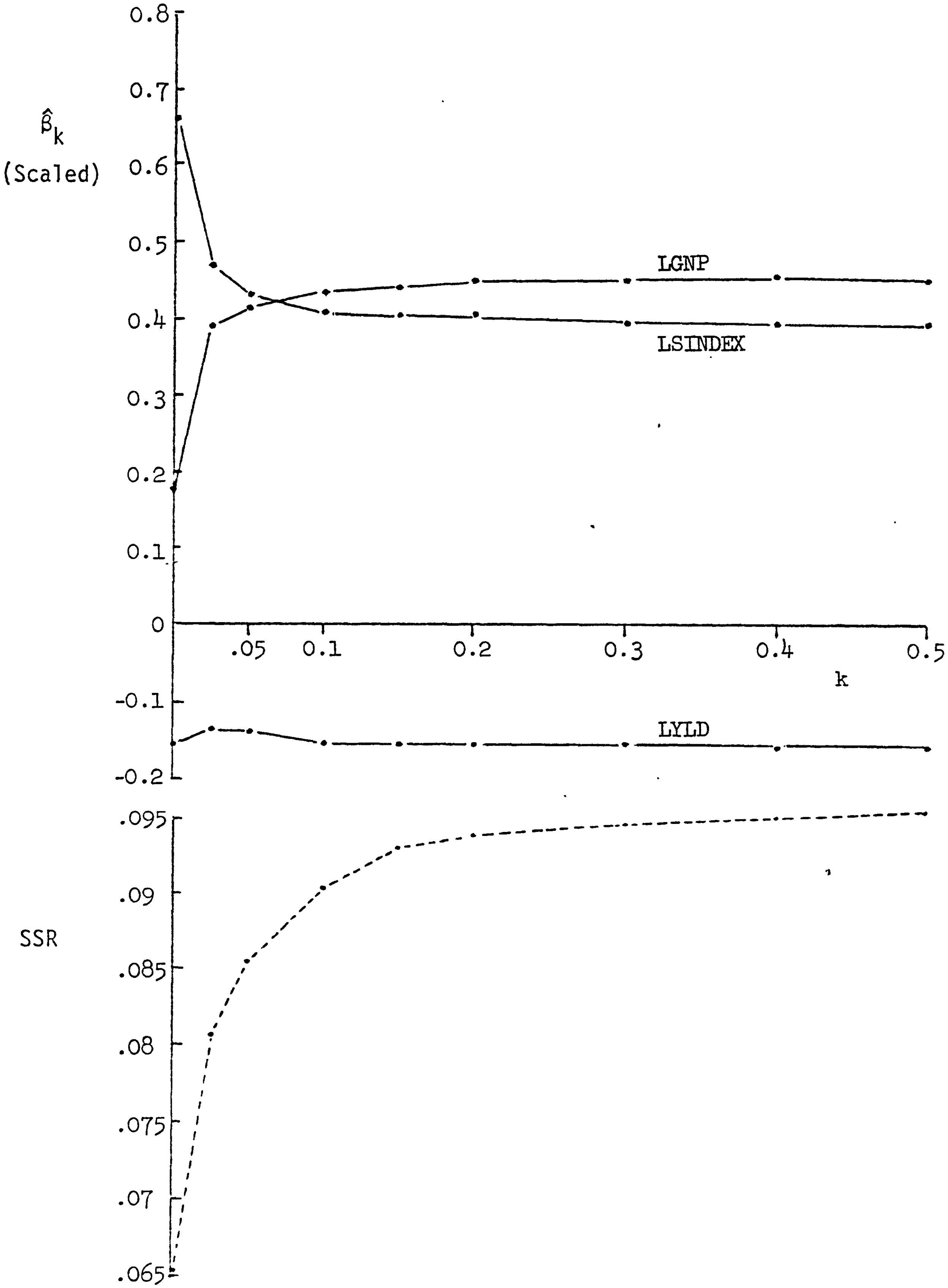
Interpreting the Ridge Trace - A Guide to Selecting the Value of k

Consider Equation (VII.4). When k=0 $\hat{\beta}_k$ is the generalized least squares (GLS) estimate. These scaled GLS estimates are shown on the Ridge Trace on the left at k=0. The two coefficients LSINDEX, the log form of service index, and LGNP, the log form of GNP, are undoubtedly inflated and unstable. Moving a short distance to the right away from the GLS estimates, these two coefficients show a rapid change in absolute value. LSINDEX, the largest coefficient, is quickly being driven down, while LGNP counteracts this movement. This is not surprising since their simple correlation coefficient of 0.9906 (Table 7) indicates that they are really the same factor with different names, and the positive sign of their covariance is driving them together. At a value of k near 0.1 the system begins to stabilize. Coefficients chosen from a value of k beyond 0.1 will undoubtedly

Figure E

RIDGE TRACE

Includes Prior Coefficients
from Generalized Least-Squares



be closer to actual values of $\hat{\beta}$, be more stable for prediction than the GLS estimates, and have the general characteristics of an orthogonal system. The tradeoff here is an increase in the standard error of the estimate with increasing values of k. The residual sum of squares (SSR) is superimposed onto the Ridge Trace to demonstrate this effect. For this reason it was decided to choose k at 0.1.

Other aids in determining reasonable choices for k are available. It may be helpful, for example, to view the least squares estimators from a Bayesian point of view. The Ridge macro in the TROLL computer software system provides the option to useone of three empirical Bayes choices of k: k_a, k_d, or k_m. k_a and k_d are both found by fitting moments of the marginal distribution of Y; while k_m is based on the marginal likelihood function of Y. The equation for k_a is easily computed:

$$k_a' = \frac{\text{Trace } (X'X)\hat{\sigma}^2}{|y-x|_w^2 - (N-2)\hat{\sigma}^2}, \qquad \text{(VII.6)}$$

whereas k_d and k_m are found by an iterative (Newton's method) solution of two equations. For this reason k_a is more appealing. Holland's paper [1973] provides a comprehensive study of these techniques.

The three empirical Bayes choices for this study are:

$$k_a = 0.11$$

$$k_d = 0.001$$

$$k_m = 0.001$$

While the k_d and k_m Bayes values result in a lower standard error, the Ridge Trace shows that the coefficients are still inflated and unstable at k equal to 0.001. so, a value of k at 0.1 remains the optimal value, and the regression coefficients at k = 0.1 are shown by the following equation:

$$\log RPM = 2.365 - 0.726 \log YLD + 1.410 \log GNP + 0.703 \log SINDEX$$

$$(10.234) \qquad (43.080) \qquad (36.252) \qquad (VII.7)$$

VIII. Robustness of the Regression Equation

Throughout the development of the air travel demand model, in its statistical evaluation, and in the correction procedures employed, the classical least-squares approach for estimating the coefficients was assumed. The justification for using this method was that the estimators of the parameters were the best linear unbiased estimators(BLUE). This result is known as the Gauss-Markov Theorem. One important realization that is contained within this theorem is that it is not necessary to make an assumption that the disturbance term follow a particular probability distribution, normal or otherwise [Murphy, 1973]. Although this assumption is not necessary, it is common and convenient to assume normality for almost all of the test procedures used. The Durbin-Watson test, for example, assumes that the disturbance term follows a normal distribution.

Suppose that the distribution of observations deviates slightly from this presumed normal distribution. Will the distribution of the estimators change only slightly? According to Huber[1973] uncontrollable variance in the error term and long-tailed error distributions will impair the efficiency of the estimates. "Just a single grossly outlying observation may spoil the least squares estimate..."[4] Hampel outlines three main causes of such deviations:

1) rounding of the observations,

2) the occurance of gross errors,

and 3) the model itself may only be an approximation to the underlying chance mechanism.

[4] Frank R. Hampel, "A General Qualitative Definition of Robustness," The Annals of Mathematical Statistics ,Volume 42, No. 6, 1971, pp. 1887-1896.

The Hat Matrix

One good form of insurance is to test the model for outliers or contaminated observations. Traditionally, an examination of the plot of the residuals (see Figure F) has been used to indicate suspect data. While such a frequency or time distribution often is informative (indicating heteroscedasticity, skewness, multiple modes, or a heavy-tailed distribution) large outliers may not be so evident since the squared error criterion heavily weights extreme values.

Hoaglin and Welsch[1977] suggest looking at the "hat matrix". The hat matrix is defined by:

$$H = X(X'X)^{-1}X' \qquad \text{(VIII.1)}$$

The diagonal elements of H, termed h_i, are interpreted as the amount of leverage or influence that the resonse value y_i exerts on the corresponding fitted value $\hat{y}_i$. The larger the size of a diagonal element, the more influence that data point exerts on the fit. Since $\sum_{i=1}^{n} h_i = p$, then the average size of the diagonal element is shown to be equal to p/n,

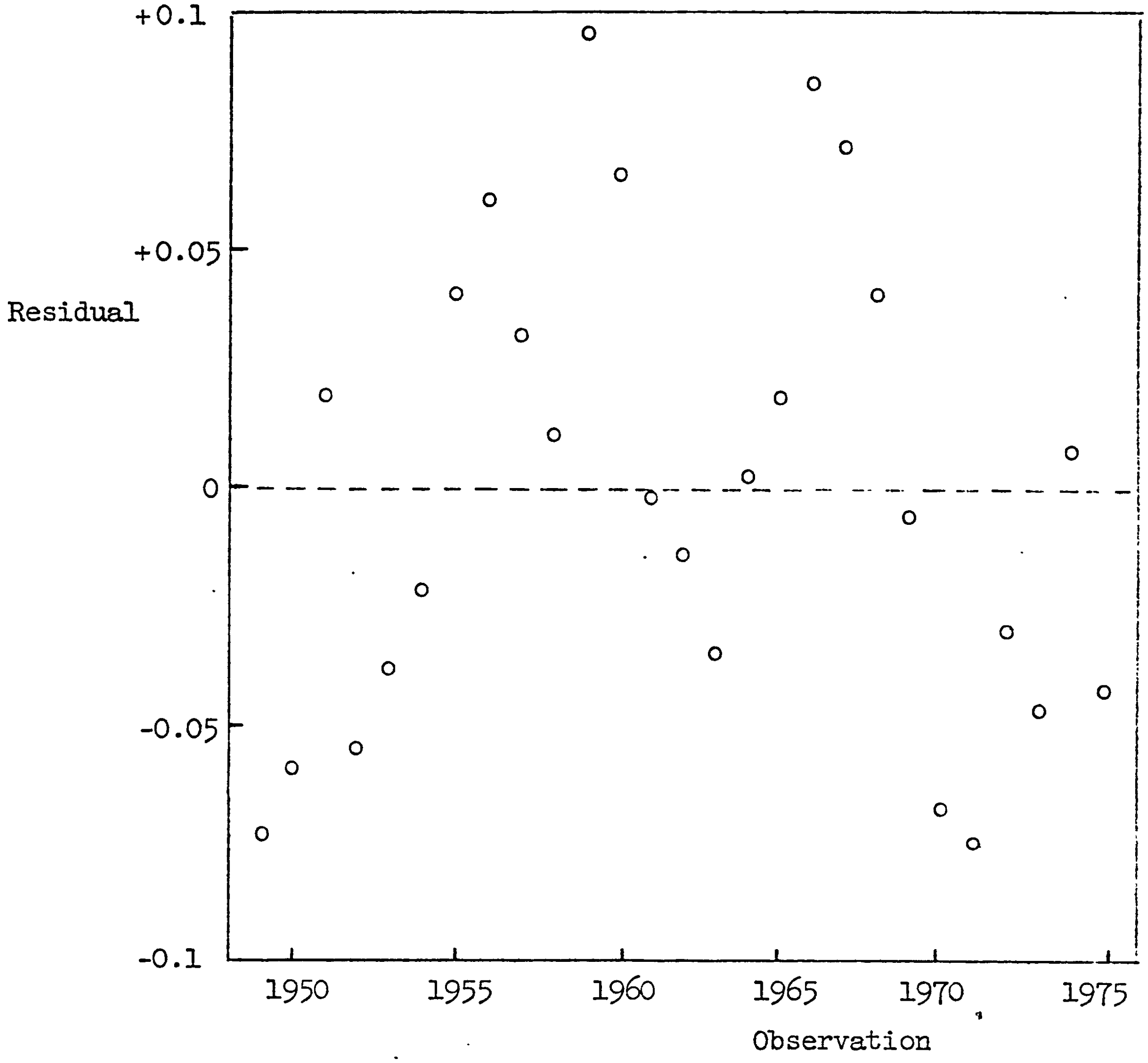
$$E[h_i] = p/n \qquad \text{where } p=1 + \text{number of explanatory variables, and } n=\text{number of observations.}$$

As a rule of thumb, Hoaglin and Welsch state that an observation "i" corresponding to the diagonal element h_i is a leverage point (outlier) if h_i is more than twice its epected value; that is, if

$$h_i > 2p/n. \qquad \text{(VIII.2)}$$

Figure F

Plot of Residuals



The calculated values of h_i for the X matrix of Equation (VII.7) are:

	h_i		h_i		h_i
1949	0.169	1958	0.189	1967	0.081
50	0.222	59	0.070	68	0.110
51	0.303	60	0.125	69	0.108
52	0.115	61	0.224	70	0.185
53	0.086	62	0.323	71	0.113
54	0.171	63	0.098	72	0.111
55	0.091	64	0.078	73	0.206
56	0.096	65	0.094	74	0.166
57	0.186	66	0.112	75	0.166

The average value of h_i is:

$$E[h_i] = p/n = 4/27 = 0.148$$

Observation "i" is a leverage point if $h_i > 2p/n = 0.296$.

Two observations are suspect - 1951 and 1962. Although neither one is truly an "extreme" outlier, they do exert more than twice the average influence on the regression fit - a fact that is not immediately evident from an examination of a plot of the residuals. In the following section a regression technique is described which is designed to make the model resistant to leverage points. Such a model is termed a "robust" model.

Robust Regression

Robust regression is a relatively new technique intended to make the regression model insensitive to changes or irregularities in the data or the assumed model. Generally, one would like to have minor inaccuracies in the model cause only small errors in the final results. Thus, a robust fit is not influenced by leverage points or contaminated observations. Robust properties are discussed in more detail in Hampel [1968 & 1971] and

Huber [1972 & 1973]. The purpose of robust regression is two-fold: first, it can be used as a diagnostic tool to determine whether leverage points unduly influence the fit,or whether the assumption of a Gaussian distribution of error terms is appropriate; and second, if necessary it can be used as a corrective procedure to provide estimates of the coefficients intended to be resistant to outliers or inaccuracies in the model.

Generally speaking, the approach is to examine the sensitivity of the regression coefficients to perturbations in the data points. The technique that will be used is called iteratively reweighted least squares. It is a modified form of Newton's method for solving the normal equations. At each iteration the residuals from the previous step are scaled, and if they are large they are given a weight which is small enough such that in the next iteration they are, in effect, deleted. If the model is "robust" the estimators will be insensitive to such perturbations. This idea will now be formalized. Consider the regression model:

$$Y = X\beta + \varepsilon \, . \qquad \text{(VIII.3)}$$

In least squares estimation the sum of the squares of the residuals is minimized - that is,

$$\min_{\beta} \sum_{i=1}^{n} (y_i - x_i \hat{\beta})^2 \qquad \text{(VIII.4)}$$

Suppose now one attempts to minimize the least <u>absolute</u> residuals, then:

$$\min_{\beta} \sum_{i=1}^{n} (y_i - x_i \hat{\beta}) \qquad \text{(VIII.5)}$$

In robust regression the residuals are weighted through some "loss function" ρ:

$$\min_{\beta} \sum_{i=1}^{n} \rho\left(\frac{y_i - x_i\hat{\beta}}{s}\right) \qquad \text{(VIII.6)}$$

where s is a measure used to scale the residuals. Differentiating Equation (VIII.6) the necessary condition for a minimum is obtained:

$$\sum_{i=1}^{n} x_{ij}\rho' \left(\frac{y_i - x_i\hat{\beta}}{s}\right) = 0 \qquad \text{(VIII.7)}$$

The way one solves this equation for $\hat{\beta}$ is by iterating - similar to Newton's method for solving the normal equations. This means that one must begin with a set of starting coefficient values $\beta°$. Residuals are computed from $\beta°$, and the weights are computed from these residuals. Let $w°$ be the vector of weights found from the starting coefficients $\beta°$; then a new set of coefficients are estimated by

$$\hat{\beta}^1 = (X'W°X)^{-1} X'W°Y \qquad \text{(VIII.8)}$$

This process is repeated. At the j^{th} step:

$$\hat{\beta}^j = (X'W^{j-1}X)^{-1} X'W^{j-1}Y \qquad \text{(VIII.9)}$$

which should converge, ordinarily within about four steps, to a local minimum of Equation (VIII.6). If one does not have a good starting guess for β, the starting coefficient estimates can be obtained from the least absolute residuals (LAR) criterion function,

$$\min \sum_{i=1}^{n} |y_i - x_i\beta| \qquad \text{(VIII.10)}$$

which is scale invariant so that one does not need the scale factor s.

For the model under investigation the ridge estimates (obtained in Section VII) can be used as starting coefficients, $\beta°$, for robust analysis. In this case when one iterates to find $\hat{\beta}$ the following would be used:

$$\hat{\beta}_k^j = \delta + (X'W^{j-1}X + kI)^{-1} X'W^{j-1}(Y - X\delta) \qquad \text{(VIII.11)}$$

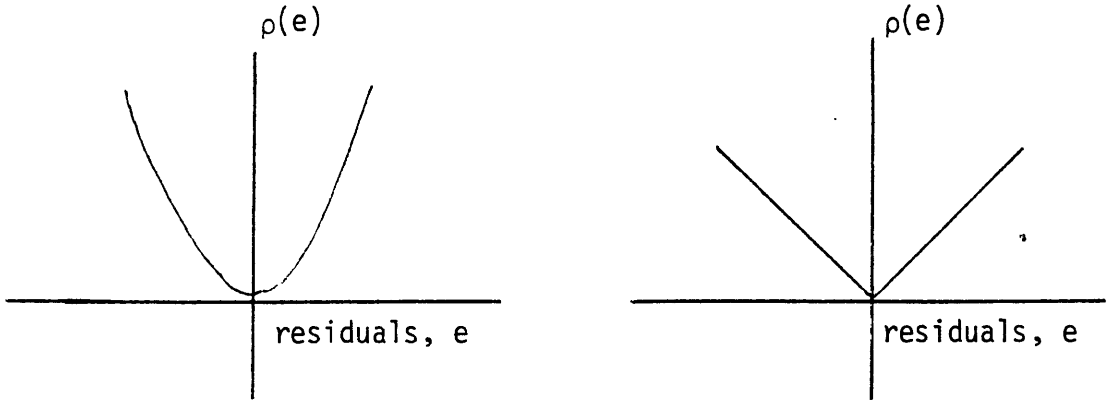
which includes the ridge parameter k and priors δ (the vector of coefficients estimated from the generalized least squares procedure in Section VI).

The vector of weights, w^j, are determined from the residuals of the previous step. They are used to "down-grade" large residuals - in effect deleting that observation from the next estimate. If $e_i^{j-1} = y_i - x_i\beta^{j-1}$ is to residual for the ith row, j-1st step, then w^j, which is a diagonal matrix with entries w_i^j, is defined as:

$$w_i^j = \frac{\rho'_e(e_i^{j-1}/s^{j-1})}{(e_i^{j-1}/s^{j-1})} \qquad \text{(VIII.12)}$$

where ρ'_e is the first derivative of the weighting or "loss" function of the residuals e.

Two possible loss functions are the least squares (LS) and the least absolute residuals (LAR)(see Figure G). For large residuals the LS function yields a more extreme weight value than the LAR function.



LS Function

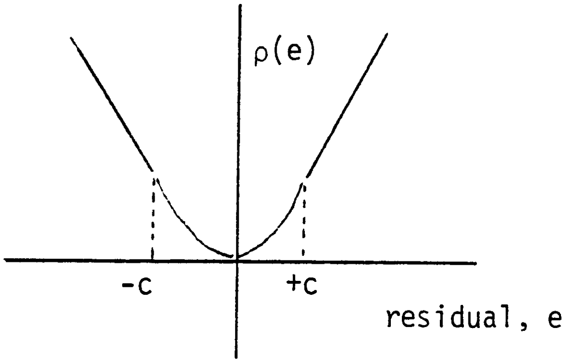
LAR Function

Figure G

Huber[1973] proposed a ρ(e) function which can be viewed as a combination of both the LS and LAR loss functions. This function is quadratic in the middle and linear for large residuals.. For example:

$$\rho(r) = \frac{1}{2}e^2 \quad \text{for } |e| < c$$
$$= c|e| - \frac{1}{2}c^2 \quad \text{for } |e| \geq c \qquad \text{(VIII.13)}$$

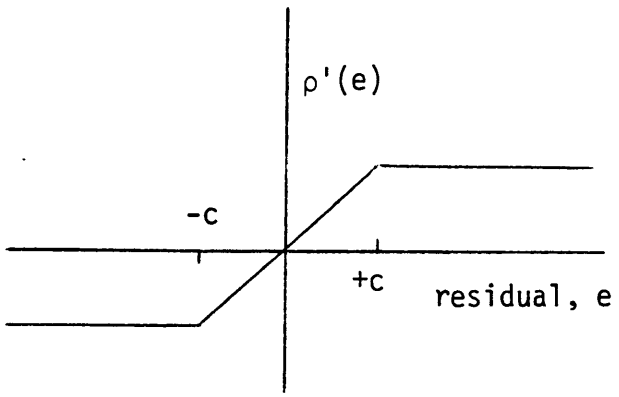
The value of c is the breakpoint between the quadratic and linear sections of the function as shown in Figure H.



The Huber Loss Function
Figure H

Note that if c=0 the Huber loss function behaves like the LAR function, and if c=∞ it behaves like the LS function.

The derivative of the Huber loss function is: ρ'(e) = max(-c,min(c,e)), which has the useful properties that: it gives normal weights to normal-sized residuals and lower weights to large residuals (thus, it replaces the least squares technique with some expression which is less sensitive to extreme values of the residuals) ; and, it is continuous and bounded (See Figure 1).



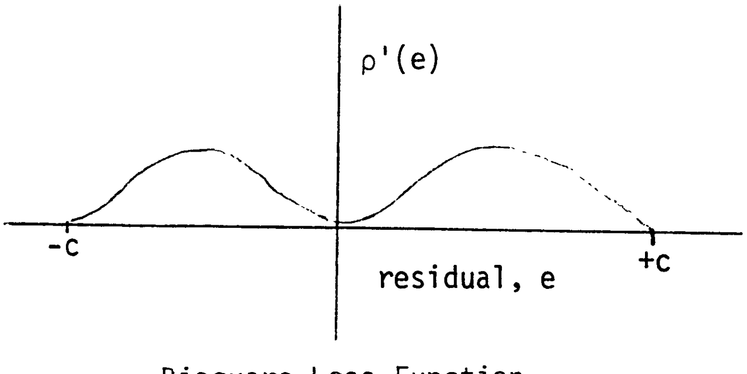
Derivative of the Huber Loss Function

Figure I

Other loss functions exist. An extreme case of the Huber loss function is a function which is redescending to zero for large residuals. One such function is the bisquare (Figure J):

$$\rho'(e) = e(1 - (e/c)^2)^2 \quad \text{for } |e| < c$$

$$= 0 \qquad |e| \geq c \qquad \text{(VIII.15)}$$



Bisquare Loss Function

Figure J

Beyond point c, observations are deleted, and thus have no influence on the estimation procedure. This is useful in determining the sensitivity of the coefficients to the removal of observations.

The computer Research Center of the National Bureau of Economic Research has developed a TROLL system macro to perform robust regression. The &ROBUST macro provides three different families of loss functions: the Huber loss function, the Bisquare loss function, and the Sine loss function (which behaves similarly to the Bisquare). All will behave like the LS function when $c = \infty$;when $c = 0$ the Huber function will behave like LAR, while the Bisquare and Sine weights will be equal to zero.

After choosing one of the three loss functions, the sensitivity of the regression fit to changes in the weights given to the residuals can be examined by plotting the coefficients against values of c. The resulting plot, called a "robust trace", will be similar to the ridge trace developed in Section VII. For convenience and similarity to the ridge trace the plot uses $r = 1/(1+c)$ as an index instead of c. Thus, when $c = \infty$, $r = 0$; and when $c = 0$, $r = 1$.

Interpretation of the Robust Trace

In Figure K is the robust trace for the problem under investigation. This trace was constructed using the TROLL &ROBUST macro, which solved equation (VIII.11) for r=1 and for each user specified value of r. Here, the Huber loss function was employed with: a ridge estimate of k=0.1, starting coefficients β° taken from ridge 0.1 (Section VII). and priors δ from the generalized least squares estimate (Section VI).

On the left-hand side of the trace, when r=0 and $c=\infty$, the Huber loss function behaves as the least squares function.[5] Any large residuals are given a normally high weighting. At the other end of the plot (on the right), when r=1 and c=0, the Huber function behaves as the least absolute residuals function. At this point large residuals are given small weights in an attempt to down-grade leverage points or contaminated observations. So at each end of the trace the residual weighting functions are very different, yet the values of the coefficients do not change significantly. Perhaps the large residuals are not down-graded enough - that is, perhaps the weights on the large residuals are too small to notice any difference. As a test, the Bisquare loss function of Equation (VIII.15) can be used which redescends to zero for large residuals. If there are any leverage points which do influence the fit, then deleting them from the regression will affect the coefficients.

Figure L is the robust trace using the Bisquare loss function. Again, at r=0 and c=∞ the Bisquare function behaves as the least squares function, so there should be no change here over the Huber analysis. But, for small values of c, r near or at 1, the Bisquare function gives large residuals negligible weights, thus setting aside substantial portions of the data. Once again, values of the scaled coefficients do not change significantly. Thus, the insensitivity of the coefficients to changes in the observations (via the weighting functions) leads to the conclusion that the regression model given by Equation (VII.7) is "robust".

[5] Note that the scaled coefficients at this point are not the ordinary least squares coefficient estimates since the estimate $\hat{\beta}$ includes the ridge parameter k as well as generalized least squares priors.

Figure K

Robust Trace with Huber Loss Function

Includes Starting Coefficients from Ridge 0.1
and Priors from Generalized Least-Squares

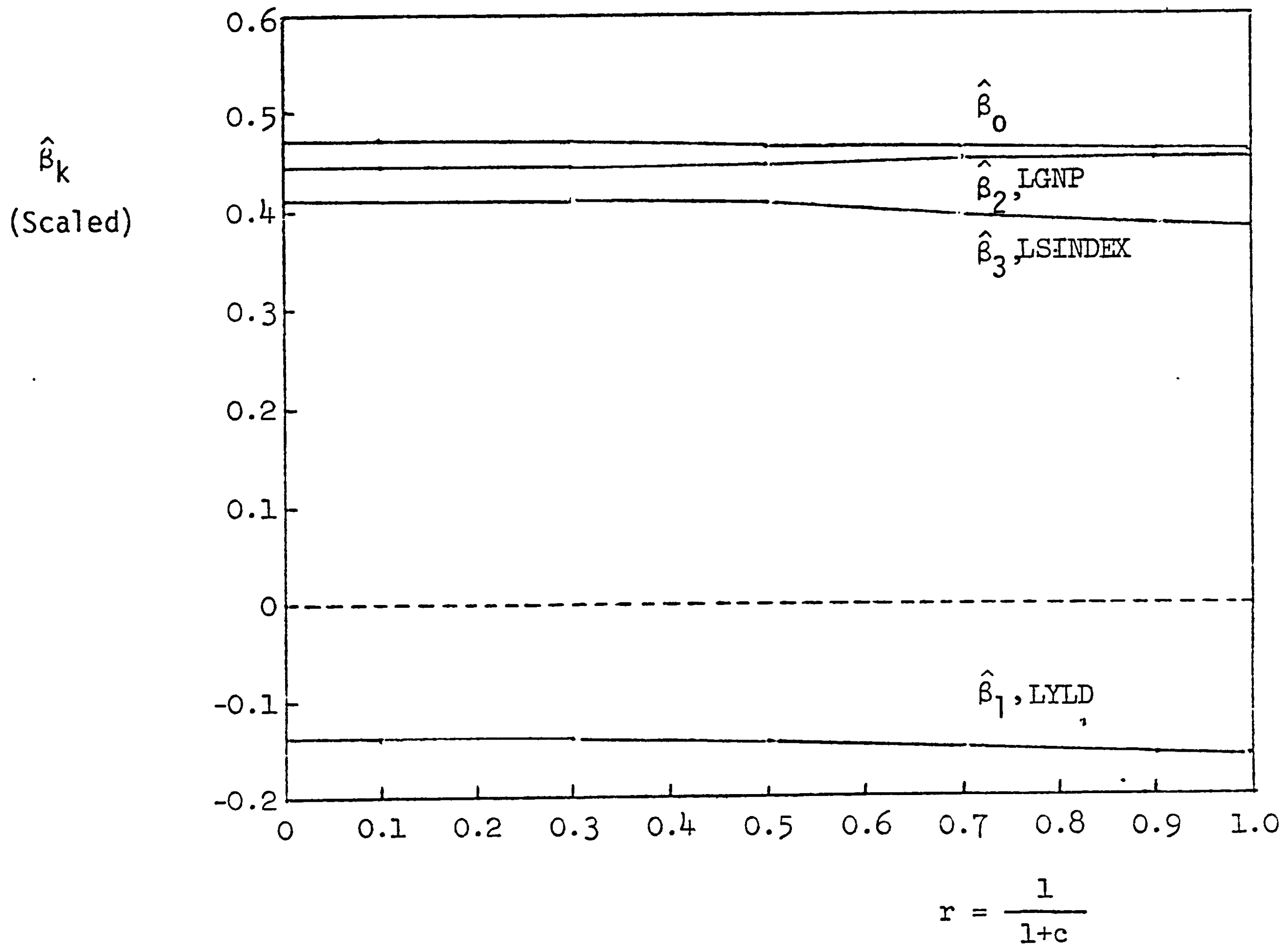
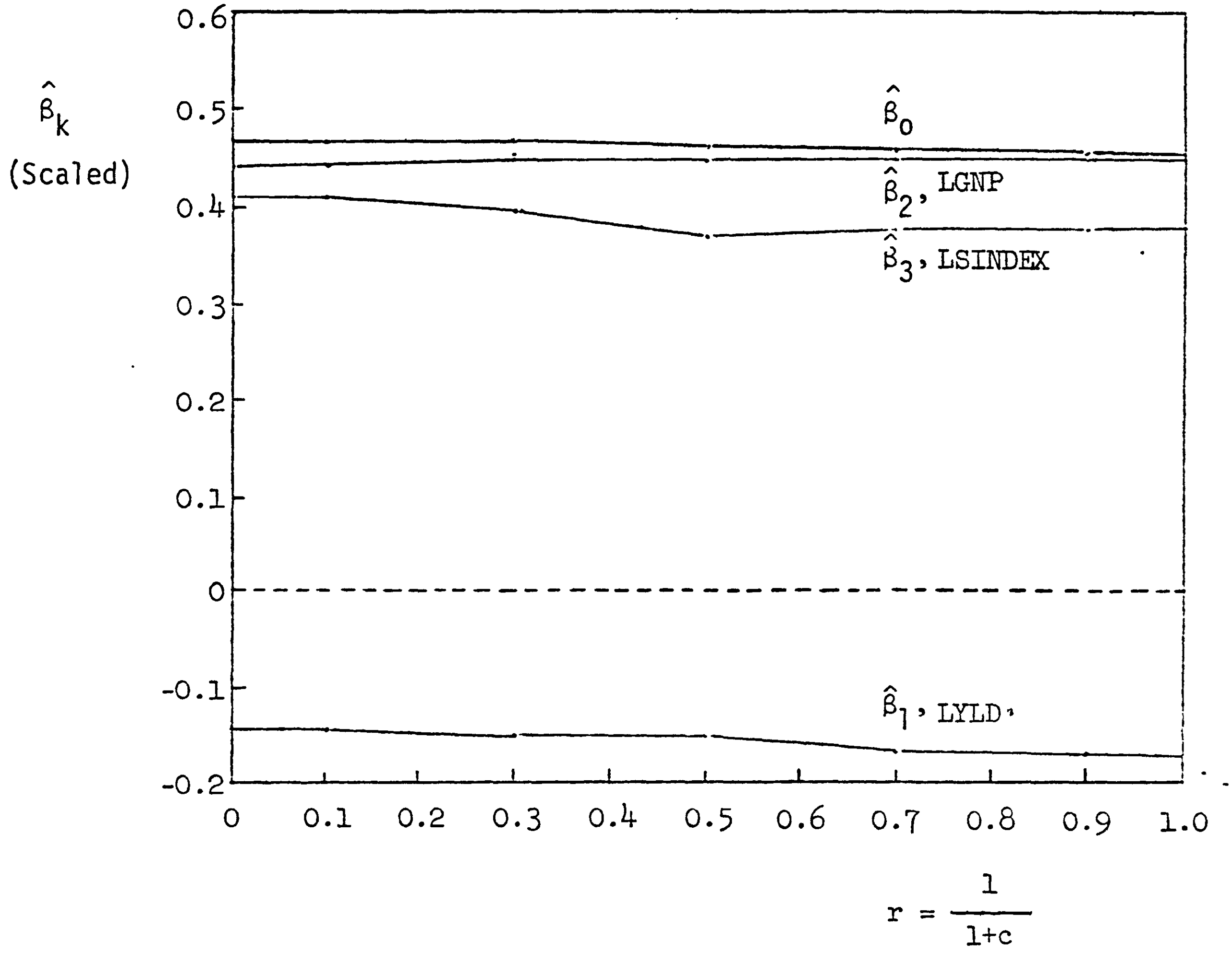


Figure L

Robust Trace with Bisquare Loss Function

Includes Starting Coefficients from Ridge 0.1
and Priors from Generalized Least-Squares



IX. Conclusions

It is evident from the preceding sections that the process of developing effective regression models is an extremely complex operation. Yet, to develop analytical models which not only provide accurate forecasts but which can also be used for policy analysis, it is essential to follow a procedure exemplified by the outline of these sections. Whereas the criteria for a "good" forecasting model might be low forecast errors, the criteria for "good" policy models is not only access to variables which are under the control of the decision maker, but also reliable knowledge about the impact of changes in these variables on the demand for air transportation. In this study an attempt has been made to demonstrate the development of a model which satisfies both of these objectives. The final model is not only a good forecasting tool, but the individual regression coefficients are statistically significant as well as "robust" for analyzing policy decisions with respect to such factors as fares, the quality of service offered, and the state of the economy.

The emphasis throughout the study has not been to develop "the ultimate model", but rather to demonstrate the use of the latest statistical estimation procedures which are not only appropriate but essential in eliminating the common problems encountered in regression models. Although some of these techniques, such as Ridge and Robust regression, are still in the experimental stages, sufficient research has been performed to warrant their application to significantly improve the currently operational regression models.

APPENDIX A

Data Values and Sources

Table 1

Data Values and Sources

Time[1]	RPM[2]	YLD[3]	CPI[4]	GNP[5]	DJLAG[6]	UNEMPLY[7]	PERATIO[8]	GOVBND[9]	INDBND[10]	SINDEX[11]	DPI[12]	IIP[13]
13	6571	6.685	71.4	387.4	181	5.9	6.49	2.31	2.74	0.671	187.1	38.8
14	7766	6.434	72.1	422.1	178	5.3	7.15	2.32	2.67	0.713	205.5	44.9
15	10211	6.482	77.8	455.4	215	3.3	8.57	2.57	2.89	0.778	224.8	48.7
16	12121	6.430	79.5	473.0	257	3.0	10.57	2.68	3.00	0.928	236.4	50.6
17	14298	6.301	80.1	490.8	271	2.9	9.77	2.93	3.30	1.022	250.7	54.8
18	16235	6.056	80.5	484.5	275	5.5	11.75	2.54	3.09	1.100	255.7	51.9
19	19206	5.913	80.2	517.2	335	4.4	12.59	2.80	3.19	1.154	273.4	58.5
20	21643	5.875	81.4	528.5	442	4.1	13.25	3.08	3.50	1.227	291.3	61.1
21	24500	5.855	84.3	538.0	491	4.3	12.73	3.47	4.12	1.357	306.9	61.9
22	24436	6.214	86.6	536.3	476	6.8	16.33	3.43	3.98	1.409	317.1	57.9
23	28127	6.458	87.3	569.7	492	5.5	17.32	4.07	4.53	1.468	336.1	64.8
24	29233	6.679	88.7	581.6	629	5.5	16.98	4.01	4.59	1.560	349.4	66.2
25	29535	6.869	89.6	596.7	618	6.7	21.68	3.90	4.54	1.667	362.9	66.7
26	33803	6.919	90.6	630.7	688	5.5	17.39	3.95	4.47	1.840	383.9	72.2
27	38720	6.343	91.7	656.4	632	5.7	18.20	4.00	4.42	1.937	402.8	76.5
28	44568	6.224	92.9	691.0	709	5.2	18.81	4.15	4.52	2.026	437.0	81.7
29	52395	6.127	94.5	731.2	833	4.5	17.92	4.21	4.61	2.183	472.2	89.8
30	64397	5.876	97.2	774.7	907	3.8	15.15	4.66	5.30	2.297	510.4	97.8
31	76008	5.668	100.0	796.3	869	3.8	17.48	4.85	5.74	2.537	544.5	100.0
32	87268	5.623	104.2	830.3	875	3.6	17.66	5.25	6.41	2.795	588.1	106.3
33	95658	5.884	109.8	852.8	906	3.5	16.48	6.10	7.25	3.101	630.4	111.1
34	95900	6.171	116.3	849.1	877	4.9	15.69	6.59	8.26	3.277	685.9	107.8
35	97618	6.576	121.3	875.2	752	5.9	18.50	5.74	7.57	3.346	742.8	109.6
36	108006	6.666	125.3	925.0	879	5.6	18.20	5.63	7.35	3.372	801.3	119.7
37	115352	6.906	133.1	974.1	947	4.9	14.22	6.30	7.60	3.465	901.7	129.8
38	117616	7.827	147.7	956.4	924	5.6	8.94	6.99	8.78	3.444	982.9	129.3

Table 1 continued...

Time[1]	RPM[2]	YLD[3]	CPI[4]	GNP[5]	DJLAG[6]	UNEMPLY[7]	PERATIO[8]	GOVBND[9]	INDBND[10]	SINDEX[11]	DPI[12]	IIP[13]
39	119400	7.953	161.2	937.5	754	8.5	14.10	6.98	9.25	3.522	1080.9	117.8

Note: All data in this table is defined for U.S. domestic trunk operations for certificated route air carriers. The data was summarized for 48 states up through 1969 and for 50 states thereafter.

Table 1 continued...

Sources:

[1] Time is expressed as 13=1949 thru 39=1975. Note: All data in this table was summarized for 48 states up through 1969 and for 50 states thereafter.

[2] Domestic trunk scheduled revenue passenger-miles. 1966 RPM's were adjusted for a major airline strike which occurred in July and August 1966. Source: Civil Aeronautics Board, Handbook of Airline Statistics. The 1973 edition contained summary yearly data up through 1972. Later data was obtained from CAB's Air Carrier Traffic Statistics, a monthly issue.

[3] Average yield is defined as the sum of passenger revenues, excess baggage revenues, and taxes divided by revenue passenger-miles. The yield will be subsequently deflated by the consumer price index to arrive at the real cost per mile to the consumer. Source: Civil Aeronautics Board, Handbook of Airline Statistics.

[4] Consumer Price Index. Source: Economic Report of the President, January 1976, Table B-17.

[5] Gross National Product. Source: Economic Report of the President, January 1976, Table B-1 adjusted to shift base year from 1972 to 1967. Expressed in 1967 dollars.

[6] Dow Jones Industrial Average lagged by one year. Computed from the quarterly highs and lows of the Dow "30" price range. Source: Wall Street Journal Library.

[7] Unemployment Ratio - a percent of civilian labour force. Source: Economic Report of the President, January 1977, Table B-27.

[8] Price to Earnings Ratio - a ratio of price index for last day of quarter to quarterly earnings(seasonally adjusted annual rate). Annual ratios are averages of quarterly data. Source: Economic Report of the President, January 1976, Table B-81. 1946-48 data based on 50 stocks.

[9] U.S. Government bond yields - percent. Source: Statistical Abstracts, U.S. Bureau of Economic Analysis, 1976, Table 801.

[10] U.S. Industrial bond yields - percent. Source: Statistical Abstracts, U.S. Bureau of Economic Analysis, 1976, Table 801.

Table 1 continued...

[11] A quality of service index obtained by principal component analysis of five service variables. See Section III of this paper.

[12] Disposable Personal Income in current dollars. Source: Economic Report of the President, January 1977, Table B-21.

[13] Index of Industrial Production. Source: Economic Report of the President, January 1977, Table B-37.

Table 2

Variables for Level of Service Index

	ASM[1]	Flight[2] Stage Length	Psgr.[3] Trip Length	Seats[4]	Speed[5]	Principal Component Scores
48 tate→1949	11.711576	168.0	448.0	34.7	178.0	-1.329
50	13.124889	162.9	460.0	37.1	180.0	-1.287
51	15.614681	174.2	466.0	39.1	183.0	-1.222
52	19.170377	186.0	499.0	42.2	189.0	-1.072
53	23.337498	192.7	512.0	45.6	196.0	-0.978
54	26.921925	198.9	517.0	49.6	204.0	-0.900
55	31.371182	206.5	519.0	51.5	208.0	-0.846
56	35.366158	214.9	534.0	52.1	210.0	-0.773
57	41.746375	227.0	562.0	53.7	214.0	-0.643
58	42.723508	233.3	567.0	55.5	219.0	-0.591
59	48.404952	231.8	575.0	58.7	223.0	-0.532
60	52.220182	228.3	583.0	65.4	235.0	-0.440
61	56.087214	226.3	589.0	72.9	252.0	-0.333
62	63.887578	241.3	601.0	79.4	274.0	-0.160
63	72.254533	251.3	602.0	83.4	286.0	-0.063
64	80.524404	261.0	605.0	86.1	296.0	0.026
65	94.787113	279.3	614.0	89.2	314.0	0.183
66	104.668839	291.0	620.0	91.2	330.0	0.297
67	133.699795	317.6	636.0	94.4	353.0	0.537
50 68	166.870750	347.5	651.0	100.8	369.0	0.795
tate→ 69	206.434271	396.4	648.0	109.8	394.0	1.101
70	213.159880	422.8	679.0	110.4	403.0	1.277
71	221.503166	427.1	681.0	115.3	405.0	1.346
72	226.621031	422.0	685.0	118.1	404.0	1.372
73[6]	244.699120	424.6	689.2	123.8	404.0	1.465
74	233.878372	424.7	683.7	127.7	401.0	1.444
75	241.282125	426.6	697.9	130.4	403.0	1.522

[1] Total available seat miles (in billions), source: Handbook of Airline Statistics, CAB, 1973, Part II, Table 14.

[2] Average Overall Flight Stage Length (miles), Source: Handbook of Airline Statistics, CAB, 1973, Part II, Table 51.

[3] Average On-line Passenger Trip Length (miles), Source: Handbook of Airline Statistics, CAB, 1973, Part II, Table 52.

[4] Average Available Seats per Aircraft, Source: Handbook of Airline Statistics, CAB 1973, Part II, Table 53.

[5] Average Overall Airborne Speed, (miles per hour), Source: Handbook of Airline Statistic, CAB, 1973, Part II, Table 54.

[6] Data from 1973-1976 obtained from Air Carrier Traffic Statistics, CAB, December 1976, p. 5.

References

Allen, David M., "The Relationship Between Variable Selection and Data Augmentation and a Method for Prediction", Technometrics, Volume 16, No. 1, February 1974. pp. 125-127.

Andrews, D.F., "A Robust Method for Multiple Linear Regression", Technometrics, Volume 16, No. 4, November 1974, pp.523-531.

Anscombe, F.J. and Tukey, John W., "The Examination and Analysis of Residuals", Technometrics, Volume 5, No. 2, May 1963, pp.141-160.

Belsley, David A., "Multicollinearity: Diagnosing Its Presence and Assessing the Potential Damage It Causes Least-Squares Estimation", National Bureau of Economic Research Working Paper No. 154. NBER Computer Research Center, Cambridge, Mass., October 1976.

Bolch, Ben W. and Huang, Cliff J., Multivariate Statistical Methods for Business and Economics. Prentice-Hall. New Jersey,1974.

Civil Aeronautics Board, "Revenue Passenger-Mile Forecast for 1976. Scheduled Domestic Operations. Domestic Trunks", Economic Evaluation Division, Bureau of Accounts and Statistics, June 1976.

Cochrane,D. and Orcutt, G.H., "Application of Least-Squares Regression to Relationships Containing Autocorrelated Error Terms", Journal of the American Statistical Association, Volume 44, March 1949, pp.32-61

Daniel, Cuthbert and Wood, Fred S., Fitting Equations to Data, Wiley-Interscience, New York, 1971.

Davis, John C., Statistics and Data Analysis in Geology, Wiley, New York, 1973.

Durbin, J. and Watson, G.S., "Testing for Serial Correlation in Least Squares Regression,I & II. Biometrica, Volumes 37 & 38, 1950-51,pp.409-428 and pp. 159-178.

Elliott, J.W., Economic Analysis for Management Decisions, Richard D. Irwin, Homewood, Illinois, 1973.

Farrar, Donald E. and Glauber, Robert R.,"Multicollinearity in Regression Analysis: The Problem Revisited", Review of Economics and Statistics, Volume 49, 1967, pp.92-107.

Glejser, H., "A New Test for Heteroscedasticity", Journal of the American Statistical Association, Volume 64, March 1969, pp.316-323.

Goldfeld, Steven M.and Quandt, Richard E., "Some Tests for Homoscedasticity", Journal of the American Statistical Association, Volume 60,June 1965, pp. 539-547.

Gorman, J.W. and Toman, R.J., "Selection of Variables for Fitting Equations to Data", Technometrics, Volume 8, No.1, February 1966,pp.27-51.

Haitovsky, Yoel, "Multicollinearity in Regression Analysis: Comment", Review of Economics and Statistics, 1969, pp. 486-489.

Hampel, Frank R., "Contributions to the Theory of Robust Estimation", Unpublished Doctoral Dissertation. Department of Statistics, University of California, Berkeley, 1968.

Hampel, Frank R., "A General Qualitative Definition of Robustness", The Annuals of Mathematical Statistics, Volume 42, No. 6, 1971, pp. 1887-1896.

Hill, Richard W. and Holland, Paul W., "A Monte Carlo Study of Two Robust Alternatives to Least Squares Regression Estimation", National Bureau of Economic Research Working Paper No. 58, NBER Computer Research Center,Cambridge, MA. , September 1974.

Hoaglin, David C. and Welsch, Roy E., "The Hat Matrix in Regression and ANOVA", Alfred P.Sloan School of Management Working Paper No. WP 901-77, Massachusetts Institute of Technology, January 1977.

Hocking, R.R., "Criteria for Selection of a Subset Regression: Which One Should Be Used?", Technometrics, Volume 14, No. 4, November 1972, pp. 967-970.

Holland, Paul W., "Weighted Ridge Regression: Combining Ridge and Robust Regression Methods", National Bureau of Economic Research Working Paper No. 11, NBER Computer Research Center,Cambridge,MA., September 1973.

Hoerl, Arthur E.,"Application of Ridge Analysis to Regression Problems", Chemical Engineering Progress, Volume 58, No. 3, March 1962,pp.54-59.

Hoerl, Arthur E., "Ridge Analysis", Chemical Engineering Progress Symposium Series, Volume 60, No. 50, 1964, pp. 67-78.

Hoerl, Arthur E.and Kennard, Robert W., "Ridge Regression: Biased Estimation for Nonorthogonal Problems", Technometrics , Volume 12, No. 1, February 1970, pp. 55-67.

Hoerl, Arthur E. and Kennard, Robert W., "Ridge Regression: Applications to Nonorthogonal Problems", Technometrics, Volume 12, No. 1, February 1970, pp. 69-82.

Huang, David S., Regression and Econometric Methods, Wiley, New York, 1970.

Huber, Peter J., "Robust Statistics: A Review", The Annals of Mathematical Statistics, Volume 43, No. 4, 1972, pp. 1041-1067.

Huber, Peter J., "Robust Regression: Asymptotics, Conjectures and Monte Carlo", The Annals of Mathematical Statistics,Volume 1, No. 5, 1973, pp. 799-821.

Johnston, J, Econometric Methods, 2nd Edition, McGraw-Hill,New York, 1972.

Kmenta, Jan, Elements of Econometrics, Macmillan, New York, 1971.

Marquardt, Donald W., "Generalized Inverses, Ridge Regression, Biased Linear Estimation, and Nonlinear Estimation", Technometrics, Volume 12, No. 3, August 1970, pp.591-611.

Marquardt, Donald W. and Snee, Ronald D., "Ridge Regression in Practice", The American Statistician, Volume 29, No. 1, February 1975, pp.3-20.

Mallows, C. L., "Choosing Variables in a Linear Regression: A Graphical Aid". Presented at the Central Regional Meeting of the Institute of Mathematical Statistics, Manhattan, Kansas, May 7-9, 1964.

Mayer, Lawrence S. and Willke, Thomas A, "On Biased Estimation in Linear Models", Technometrics, Volume 15, No. 3, August 1973, pp.497-508.

Miller, Leslie H., "Table of Percentage Points of Kolmogorov Statistics", Journal of the American Statistical Association, March 1956, pp.111-121.

Murphy, James L., Introductory Econometrics, Richard D. Irwin, Homewood, Ill., 1973.

Neter, John and Wasserman, William, Applied Linear Statistical Models, Richard D. Irwin, Homewood, Ill., 1974.

Prescott, P., "An Approximate Test for Outliers in Linear Models", Technometrics, Volume 17, No. 1, February 1975, pp.129-132.

Taneja, Nawal K., "Statistical Evaluation of Econometric Air Travel Demand Models", Journal of Aircraft, Volume 13, No. 9, September 1976, pp. 662-669.

Wonnacott, R. J. and T. H. Wonnacott, Econometrics, Wiley, New York, 1970.