A PILOT STUDY OF THE LUNG FUNCTION EFFECTS IN HEALTHY ADULTS DUE TO ACUTE EXPOSURE TO AMBIENT AIR POLLUTION IN PORTAGE, WISCONSIN IN 1978

BY

ROBERT JOHN ROSCOE

A thesis submitted in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE
(Preventive Medicine-Epidemiology)

at the

UNIVERSITY OF WISCONSIN—MADISON

1980

A PILOT STUDY OF THE LUNG FUNCTION EFFECTS IN HEALTHY ADULTS DUE TO ACUTE EXPOSURE TO AMBIENT AIR POLLUTION IN PORTAGE, WISCONSIN IN 1978

Abstract

As indicators of pulmonary function, the forced vital capacity (FVC) and forced expiratory volume in one second ($FEV_{1.0}$) were evaluated from spirometric traces taken in the autumn of 1978 in Portage, Wisconsin on four healthy adult females. The time series of FVC and $FEV_{1.0}$ observations were regressed on simultaneous time series of obervations of sulfur dioxide, nitrogen dioxide, ozone, air temperature and dew point temperature. All possible subsets and principal components regressions were used and each subject was treated individually as a separate experiment. No regressions were judged significant based on a preplanned validation criterion. More importantly, the lung function data were judged to be invalid based on the presence of a large "fatigue effect" which produced decreasing trends in the response variable. Among other problems identified were confounding due to low readings on one of four spirometers and extreme case deletion due to missing data. Potential solutions were proposed. A methodology based on the findings of this pilot study was proposed to be applied to superior data from Portage and five other cities of varying pollution levels.

1. Introduction

This research was designed to facilitate a "study within a study" which is intended to contribute to the body of information on which revisions of current 24-hour (h) air quality standards for sulfur dioxide and particulates and 1-h standards for oxidants will be considered. In addition, this acute effects substudy is designed to gather data on suspended sulfates, mass respirable particulates and nitrogen dioxide which might aid decision-makers in their determination of whether short-term standards for these pollutants are necessary. The main study under which the acute effects substudy falls is the "Six Cities Lung Health Study" currently being conducted by the Harvard University School of Public Health (Speizer et al.: 1977 and Ferris et al.: 1978a). The results of this prospective study of the chronic effects of air pollutants on lung function of adults and children in varied environments will likely contribute to the revision or maintenance of current annual air quality standards. The "study within a study" arises from the daily lung function testing that was performed by the Harvard interviewers upon themselves during their stays in each of the six cities to complete the annual population lung function testing for the main study. Thus the interviewers in the main study become the subjects in in the substudy. This daily spirometric lung function testing was specifically requested by the U.S. Environmental Protection Agency.

The acute effects substudy will consist of the quantification and comparison of lung function effects due to acute exposure to air pollutants based on daily spirometric data from each of the six cities. The research presented herein describes the formulation of a methodology based on a pilot study of daily interviewer spirometer data collected in one of the "six cities", Portage, Wisconsin. It is anticipated that this methodology will be applied to the daily spirometer data from each of the six cities during the implementation of the acute effects substudy.

The pilot study was designed to accomplish objectives two and three, below, in the process of achieving objective one. These objectives are:

1) To quantify any acute lung function effects from air pollutants or meteorological factors based on the limited Portage data currently available.

2) To ascertain which variables and regression procedures are appropriate for accomplishing objective one

3) To discover and control any confounding factors or other problems of significance.

Studies relating acute lung function effects to air pollutants have been well documented by National Academy of Sciences publications (1977a, 1977b, 1978 and 1979) and by Ferris (1978). Precedents for the inclusion of meteorological variables in studies relating health indicators to air pollution have been provided by Lebowitz et al. (1972) and by Hosein et al. (1977). The use of spirometric measurements for the assessment of acute health effects related to air pollutants has frequently been reported (Lawther et al.: 1974 and 1975, Linn et al.: 1976, Kerr et al.: 1975). Standardization of spirometric procedures and development of summary statistics appropriate for epidemiologic studies have been addressed by Ferris et al. (1978b and 1978c) and by Tager et al. (1976). Methodologies suitable for the study of possible relationships between acute health effects and environmental factors have been suggested by various authors. Stebbings (1978) reported on the use of the differences in day-to-day variable levels rather than the raw variable levels in panel studies of acute health effects. Shortcomings in regression procedures in studies of this type were discussed by Goldstein and Rausch (1978). Korn and Whittemore (1979) stressed the advantages of using subjects separately as individual experiments. Munn (1970) provided a broad background for undertaking studies related to the atmospheric environment. Background for the statistical procedures used in this research was obtained from Levenstein et al. (1979), Hocking (1976), Draper & Smith (1966) and Chatterjee and Price (1977).

2. Materials and Methods

2.1 Subjects

Details concerning the subjects available for this research are presented in Table 1. These subjects were members of the Harvard interviewer teams that surveyed the children's cohort in Portage in 1977 and 1978. Out of the eleven time series of daily spirometric measurements on seven subjects, four time series on four subjects were chosen for analysis. The 1978 time series for subjects 04, 80, 56 and 57 were chosen on the basis of their length. Other time series were considered too short for analysis especially when missing values among the independent variables caused more cases to be eliminated. The tests were conducted between 10:00AM & 2:00PM with 82% at either 11:00 AM or 12 noon.

The four subjects were all female and all residents of the Portage area, not travelling interviewers. Their ages varied from 29 to 64 and none had ever smoked. Respiratory health questionnaires (modified version of British MRC) completed by each subject revealed no important ~~complicating~~ confounding factors.

2.2 Lung Function Testing Protocol

The protocol for the lung function testing was the same as that used for adult subjects in the "Six Cities Lung Health Study" (Ferris et al: 1978a, 1978b & 1978c). Lung function for each subject was expressed in terms of functions of forced vital capacity (FVC) and forced

TABLE 1 SUMMARY OF SUBJECT CHARACTERISTICS

(Denotes time series chosen for analysis.)**

(L)-Local Interviewer
(T)-Traveling Interviewer

Subject Number	Number of Readings 1977	Number of Readings 1978	Sex	Age*	Height cm(in)	Weight kg(lbs)	Possible Occupat. Exposure[1]	Smoking History[1]	Respiratory Health Symptoms[1]
04 (L)	2	28 **	F	64	157(62)	57.6(127)	none	never	phlegm, high blood pressure
13 (T)	12	7	F	33	160(62)	49.0(108)	none	never	wheezing, chest infection, hay fever, allergy
22 (T)	13	16	M	26	182(70)	63.5(140)	none	never	chest cold, sinus trouble, headaches
80 (L)	0	24 **	F	41	173(67)	72.6(160)	none	never	sinus trouble
57 (L)	0	26 **	F	29	171(66)	52.6(116)	cigarette smoke, fabric dust	never	headaches
56 (L)	14	23 **	F	53	155(61)	51.3(113)	cleaning materials	never	high blood pressure
82 (L)	6	0	F	42	157(62)	45.4(100)	NO_x	never	allergies

10/3-10/25 9/15-10/25 *as of test date

[1] (from resp. health questionnaire)

41 days

high propensity

expiratory volume in one second ($FEV_{1.0}$). The $FEV_{1.0}$ and FVC readings were machine read on a digitizer from spirometric traces made on Collins Survey spirometers Stead-Wells type, ~~spirometers~~ both portable and table top models. The subjects used no nose clips. The interviewer watched for nasal leakage or other factors which might have made the trial erroneous and so indicated when necessary. Each subject received a minimum of five attempts to produce three acceptable tracings. If three acceptable tracings were not produced from five trials, further trials up to a grand total of eight were permitted. The acceptable tracings were used to select maximums and calculate means. The three readings used to calculate a mean FVC or $FEV_{1.0}$ were required to be within 200ml of each other in order to be acceptable for averaging. Variation larger than 200ml was thought to be more likely due to variability in subject effort or measurement error than it was to actual variation in FVC or $FEV_{1.0}$. If the variation among the three largest readings was larger than 200ml, the average of the maximum and second highest was used unless the range of the two exceeded 200ml in which case ~~or~~ the average of the second and third highest was used.

how?

Rationale

The temperature of the spirometer at the time of each reading was recorded. All FVC and $FEV_{1.0}$ readings were corrected to body temperature and pressure saturated with water vapor (BTPS). The BTPS corrections were calculated using equations and tables provided by Cotes (1975). Standard barometric pressure of 1013mbs (760mmHg) was used in all cases.

2.3 Dependent variables

The dependent variables under consideration for this research were the maximum and the mean of the best three out of five FVC, $FEV_{1.0}$, FVC + $FEV_{1.0}$ and $FEV_{1.0}$/FVC. The object in choosing dependent variables was to select variables which are sufficiently ~~stable~~ reproducible for an epidemiologic study yet sufficiently sensitive to measure variations in lung function due to the acute effects of air pollutants.

The ratio of $FEV_{1.0}$/FVC was eliminated on the basis of not being ~~suitable~~ physiologically correct for measuring the ~~physiologic~~ changes associated with short-term exposure to air pollutants. The sum of FVC + $FEV_{1.0}$ was eliminated on the basis of its not being significantly different from the individual measurements of FVC and $FEV_{1.0}$. The sum was highly correlated with FVC and $FEV_{1.0}$ and had approximately the same coefficient of variation. In the future the maximal mid-expiratory flow (MMEF) will be considered and in this case the MMEF will be calculated from the spirometric trace having the maximum FVC + $FEV_{1.0}$. The maximum $FEV_{1.0}$ and FVC were eliminated on the grounds of not being significantly different from the mean of the best three out of five $FEV_{1.0}$ and FVC.

Thus the dependent variables chosen for analysis were the mean of the best three out of five $FEV_{1.0}$ (FEVMN) and FVC (FVCMN). These were the primary dependent variables chosen for the "Six Cities Lung Health Study" and no compelling reason was found to substitute other variables in the substudy of acute lung function effects.

Perhaps the most important statistical consideration with respect to dependent variables in regression analysis is autocorrelation. It had been planned from the beginning to standardize the four time series in order to obtain comparable results. This was intended to control for autocorrelation due to lung size, especially when the data from the four subjects were pooled. Other possible sources of autocorrelation such as a "learning effect" which might result in a rapid increase at the beginning of a time series or a "fatigue effect" which might result in a decreasing trend were not expected. Adult lung function over a period of one to three months was not expected to decrease markedly. A mean, calculated from the readings which comprise the time series, was expected to be stationary ~~meaningful throughout~~.

These expectations were not realized. Appendix A, pp. A1-A4, contains graphs of the time series for each individual subject's FEVMN and FVCMN. In these graphs

"N" indicates the number of cases and "COR" indicates the correlation between the sequence number of the test (COUNT) and the raw lung function measurement in units of liters. Sequence number was used instead of date because weekends and occasional weekdays were missing. From these graphs it was evident that a decreasing trend was often present and sometimes striking (e.g., FVCMN and FEVMN for subject 80). These decreasing trends were considered to be too extreme to be genuine reflections of lung function. Upon consultation with interviewers who were present during these tests it was concluded that these decreasing trends were most likely the result of a subject "fatigue effect". This finding casted doubt upon the validity of the entire data set for 1978 but analysis continued for the sake of finding other problems and developing a complete methodology.

The decision taken in order to control for this autocorrelation due to "fatigue effect" was to standardize the lung function measurements for each subject using deviations from a regression line and the square root of the residual mean square error rather than deviations from a mean and the standard deviation. Thus the individual time series were regressed on sequence number using BMDP6D and the residuals from this regression were standardized as stated above. The results of this standardization for

FEVMN are presented in Figure 1, p. 11. The sequence number in Figure 1 is fixed on the time series of subject 04. (N04Y78 denotes number 04 in the year 1978, etc.) Subject 04 completed 28 tests and missed no weekdays except one holiday during that period. The missing values in the time series for subjects 80, 56 and 57 indicate days on which subject 04 was tested but others were not.

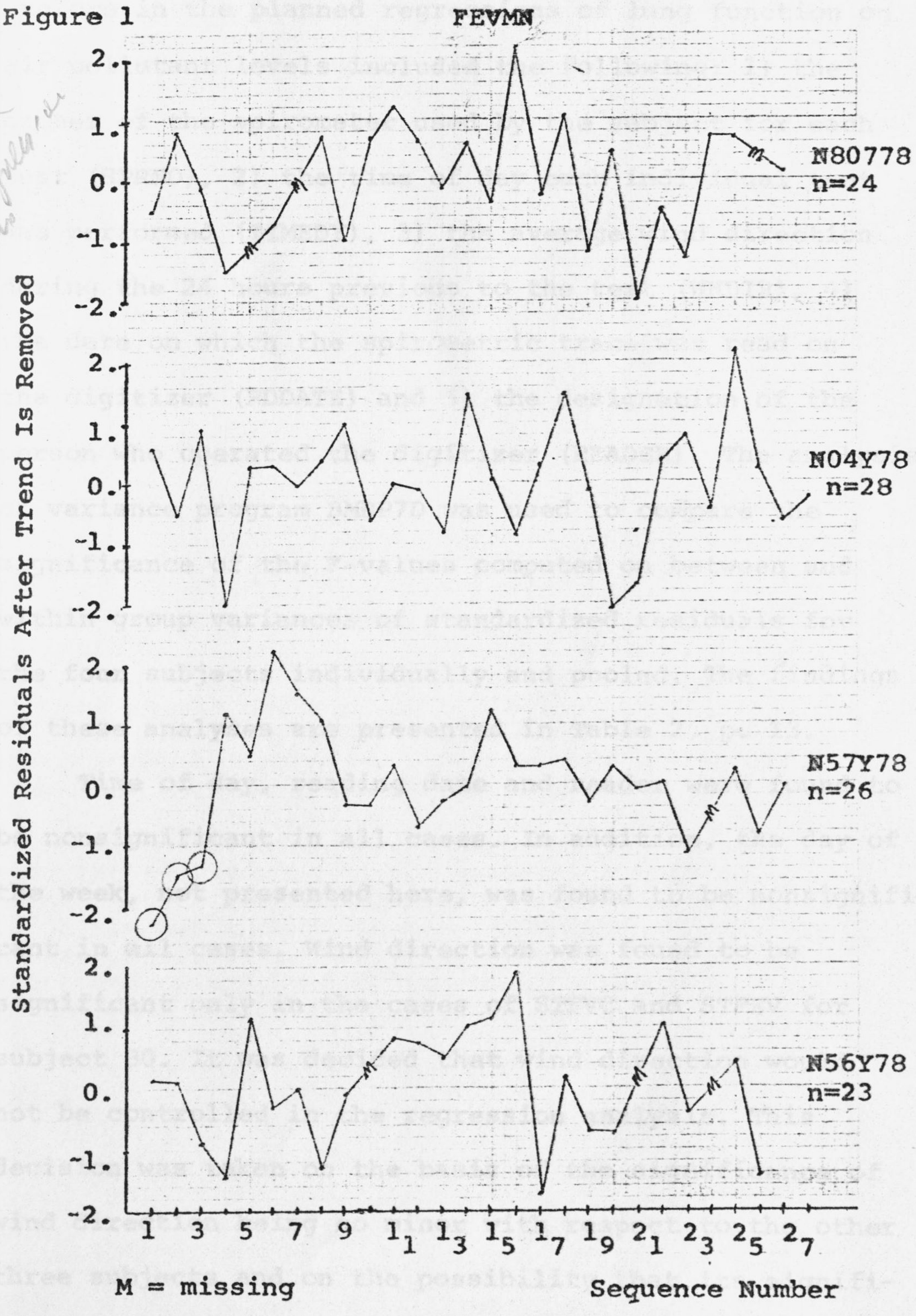
The appearance of the time series now seemed free of autocorrelation ~~consistent with statistical independence~~ except in the case of subject 57 (N57Y78). In this case the first three readings which are circled in Figure 1, seemed to be due to a "learning effect". It was decided to drop these three cases as outliers and restandardize. In Appendix A, p. A5, is a graph of the time series for subject 57 with the first three cases removed. Appendix A, p. A6, contains a graph of the standardized residuals for subject 57 with the outliers removed. This time series, too, appeared free from autocorrelation when the outliers were removed.

Appendix A, pp. A7-A9, contains graphs of the other three subjects' standardized time series. STFEV denotes standardized FEVMN and STFVC denotes standardized FVCMN.

2.4 Bias Testing

Categorical variables which were potential biasing

Figure 1


FEVMN
Standardized Residuals After Trend Is Removed
2.
1.
0.
-1.
-2.
N80778
n=24
N04Y78
n=28
N57Y78
n=26
N56Y78
n=23
1
3
5
7
9
11
13
15
17
19
21
23
25
27
M = missing
Sequence Number

factors in the planned regressions of lung function on air pollutant levels included the following: 1) the number of the spirometer used by the subject for each test (SPRNO), 2) the time of day each individual test was performed (TIMEDY), 3) the average wind direction during the 24 hours previous to the test (WNDIR), 4) the date on which the spirometric trace was read on the digitizer (RDDATE) and 5) the designation of the person who operated the digitizer (READER). The analysis of variance program BMDP7D was used to compare the significance of the F-values computed on between and within group variances of standardized residuals for the four subjects individually and pooled. The findings of these analyses are presented in Table 2, p. 13.

Time of day, reading date and reader were found to be nonsignificant in all cases. In addition, the day of the week, not presented here, was found to be nonsignificant in all cases. Wind direction was found to be significant only in the cases of STFVC and STFEV for subject 80. It was decided that wind direction would not be controlled in the regression analysis. This decision was taken on the basis of the significance of wind direction being so minor with respect to the other three subjects and on the possibility that its significance with respect to subject 80 could be an artifact of

BIAS INVESTIGATIONS OF INDIVIDUAL AND POOLED STANDARDIZED RESIDUALS*

Subject	SPRNO	TIMEDY	RDDATE	WNDIR	READER
		STFEV			
N56Y78	.0832	.9280	.9243	.7012	.9243
N57Y78	.7341	.8103	.8645	.8123	one rdr.
N04Y78	.0072	.9559	one date	.7823	one rdr.
N80Y78	.0353	.5686	.2937	.0444	.2937
Pooled	.0001	.9344	.3940	.1540	.5669
		STFVC			
N56Y78	.0385	.7899	.6076	.8407	.6076
N57Y78	.3841	.2183	.3203	.7174	one rdr.
N04Y78	.0028	.9803	one date	.6345	one rdr.
N80Y78	.0065	.8528	.2194	.0246	.2194
Pooled	.0000	.4970	.3887	.1214	.5971

Numbers in the body of the table are the tail probabilities for the F-value of the analysis of variance test of within group and between group variances. Probabilities less than .05 are considered significant.

* The standardized residuals are obtained by regressing the time series on the sequence number of the test. This is done to eliminate trends in the data which are not thought to be attributable to environmental factors. The residuals from the regression are standardized by dividing them by the square root of the residual mean square error.

TABLE 2

small sample size.

Spirometer number was found to be a significant biasing factor in all cases except for the STFEV and STFVC of subject 57 (outliers removed). In this case the possibility was considered that the results of subject 57's analysis might be erroneous due to small sample size. Validation from repeated similar results was the most important criterion for determining the probable trueness of relationships. Further investigations of the bias due to spirometer used are presented in Table 3, p. 15. Except for the case of STFEV for subject 57, spirometer number 15 produced the lowest mean results for STFEV and STFVC both individually and pooled. To make certain that the between group differences were significant due to the low readings of spirometer 15 alone, the analysis of variance test was recomputed omitting all tests performed on spirometer number 15. In every case the spirometer number was found to be nonsignificant. The decision was made to omit tests performed on spirometer 15 in the regression analysis.

2.5 Setting

The site designated "Portage" in the "Six Cities Lung Health Study" consists of a large portion of central Columbia county, Wisconsin. The 1978 population of the study area was approximately 30,000 while the City of

SPIROMETER BIAS INVESTIGATION OF POOLED AND INDIVIDUAL STANDARDIZED RESIDUALS OF FEVMN AND FVCMN

STFEV		Spirometer Number				
Subj.	Stat.	3	5	8	15	Total
56	X	.993	.210	.138	-.598	-.000
	s	.942	.424	1.078	.599	.977
	n	3	2	10	8	23
57	X	.244	.024	-.387	.445	.000 [1]
	s	1.904	.897	.665	.978	.976
	n	3	13	5	2	23
04	X	.770	-.080	.020	-.690	.000
	s	.687	.085	.773	.974	.982
	n	9	2	7	10	28
80	X	.331	-.790	.444	-.792	-.000
	s	1.035	.789	.566	1.032	.979
	n	7	3	9	5	24
Pld.	X	.550	-.015	.138	-.616	-.000 [2]
	s	1.088	.595	.824	.923	.965
	n	23	20	31	27	101
STFVC						
56	X	.893	.155	.277	-.720	.000
	s	.986	1.280	.862	.695	.977
	n	3	2	10	8	23
57	X	.244	.177	-.179	-1.071	-.000 [1]
	s	1.075	.969	1.009	.564	.977
	n	3	13	5	2	23
04	X	.806	-.045	.039	-.743	-.000
	s	.888	.445	.815	.678	.981
	n	9	2	7	10	28
80	X	.181	-.420	.601	-1.086	-.000
	s	.570	.910	.669	1.106	.978
	n	7	3	9	5	24
Pld.	X	.540	.107	.249	-.827	-.000 [2]
	s	.840	.883	.831	.749	.964
	n	23	20	31	27	101

X = mean, s = standard deviation, n = number of cases

1 Statistics for subject 57 were calculated with three outliers removed.

2 Pooled statistics were calculated with the three outliers for subject 57 included.

TABLE 3

Portage had 7738 inhabitants. The area is a rural, mainly agricultural, community specializing in dairy, alfalfa, corn, wheat and soybeans. The elevation is between 152m (500 feet) and 305m (1000 feet) with the higher elevations located primarily southwest of the City of Portage. The soils consist of loams which have formed from loess and limy glacial tills and areas of wetland soils around the Wisconsin River. The topography is flat to gently rolling. The bedrock consists of Upper Cambrian formations of sandstones with some dolomite and shale together with some Precambrian quartzite, slate and iron formations. The vastly altered natural vegetation consisted of oak-basswood forest.

Potential sources of air pollutants for the Portage area include the following: 1) particulates including allergenic substances arising from field crops, pesticides and fertilizers, 2) NO_x, O_3 and particulates arising from traffic in Portage and from traffic on Interstate Highway 90 and 94 and U.S. Highways 51 and 16, 3) O_3, particulates, total oxidants and sulfates imported from the Chicago/Gary area (250km southeast), the Milwaukee area (130km east-southeast) and the Minneapolis/St Paul area (350km northwest), 4) SO_2, NO_x and particulates arising from the Columbia Generating Station (CGS) located 6.5km south-south~~we~~ea st of Portage and in the center of the study area. CGS is a coal-fired electrical generating station which had 1977 and 1978 output capacities of 527 megawatts and 1054 megawatts respectively. Pollutants are emitted from this plant from two stacks, one at 30m and one at 91m.

Annual pollution levels for the study area for the period May, 1977 through April, 1978 estimated from a network of 12 monitoring sites located throughout the area are: 1) sulfur dioxide, mean = 8μg/m^3, range = 4-10μg/m^3; 2) nitrogen dioxide, mean = 12μg/m^3, range = 6-22μg/m^3; 3) suspended sulfates, mean = 4μg/m^3, range = 2-4μg/m^3; 4) total supended particulates, mean = 37μg/m^3. All levels are corrected to standard temperature and pressure (25C, 1013mbs). No data on ozone were available from this monitoring network.

2.6 Pollution and Meteorological Monitoring

The daily pollution data which were used in this research were collected from one site, Genrich, located 3km northeast of the City of Portage. Genrich was chosen because it was the only site at which SO_2, NO_2 and O_3 were all monitored and because it was the site located closest to the City of Portage. Most, but not all, of the daily interviewer lung function testing was done in the City of Portage. Thus Genrich offered a crude estimate of pollution exposure, however, this was the best estimate available for 1978. In the future, pollution monitoring at the site of the local interviewers' office is planned.

Hourly averages of ozone and sulfur dioxide were collected at the Genrich site by Dames and Moore, Inc., Park Ridge, Ill., under contract with Wisconsin Power and Light Company. Ozone was analyzed using a Meloy OA-350 ozone analyzer which was calibrated quarterly using an ultra-violet lamp. The detection limit for ozone was $2\mu g/m^3$. Sulfur dioxide was analyzed using a Meloy SA-285 SO_2 analyzer which was calibrated quarterly with a Monitor Lab 8500 calibrator. (SO_2 detection limit was $< 10\mu g/m^3$)/ The air intake for both O_3 and SO_2 samples was located 4.6m above the ground. Sulfur dioxide and ozone data were recorded on a Esterline-Angus L11015 single channel speed Servo II Recorder and averaged with a Monitor Lab 8640 Averager.

Nitrogen dioxide was monitored on an hourly basis at the Genrich site by a group under the supervision of Dr. Ted Tibbitts of the Department of Horticulture of the University of Wisconsin, Madison. NO_x was analyzed using a Monitor Lab 8440 Analyzer which was calibrated at two month intervals according to the Federal Reference Method using a Bendix 8861D Gas Phase Titrator modified to E.P.A. guidelines for gas titration units. The detection limit was $4\mu g/m^3$.

Total suspended particulates were collected over the 1978 Portage test period, 9/15/78-10/25/78, in the form of 24-h integrated samples every sixth day. This frequency was considered insufficient for use in the acute study of interviewer lung function. Daily particulates, including mass respirable particulates and suspended sulfates are among the pollutants designated for monitoring at the local interviewer offices in the future.

inadequate measurements for acute changes

Figure 2, p. 20, provides a display of the pollutant levels which were measured at Genrich during the 1978 test period.

Several meteorological variables were originally intended for inclusion in the analysis. Small sample sizes, however, have necessitated reducing the meteorological variables to two. One representation of temperature and one of moisture (dew point temperature), which have been associated with variations in lung function, were chosen. Pressure, precipitation, windspeed and net radiation although obviously associated with pollution levels have been eliminated on the grounds that the pollution level is of crucial importance and the factors that caused this level to occur are of secondary importance.

It was originally planned to monitor meteorological variables at sites in the Portage area such as Tower site. This was found to be inconsistent with monitoring capabil-

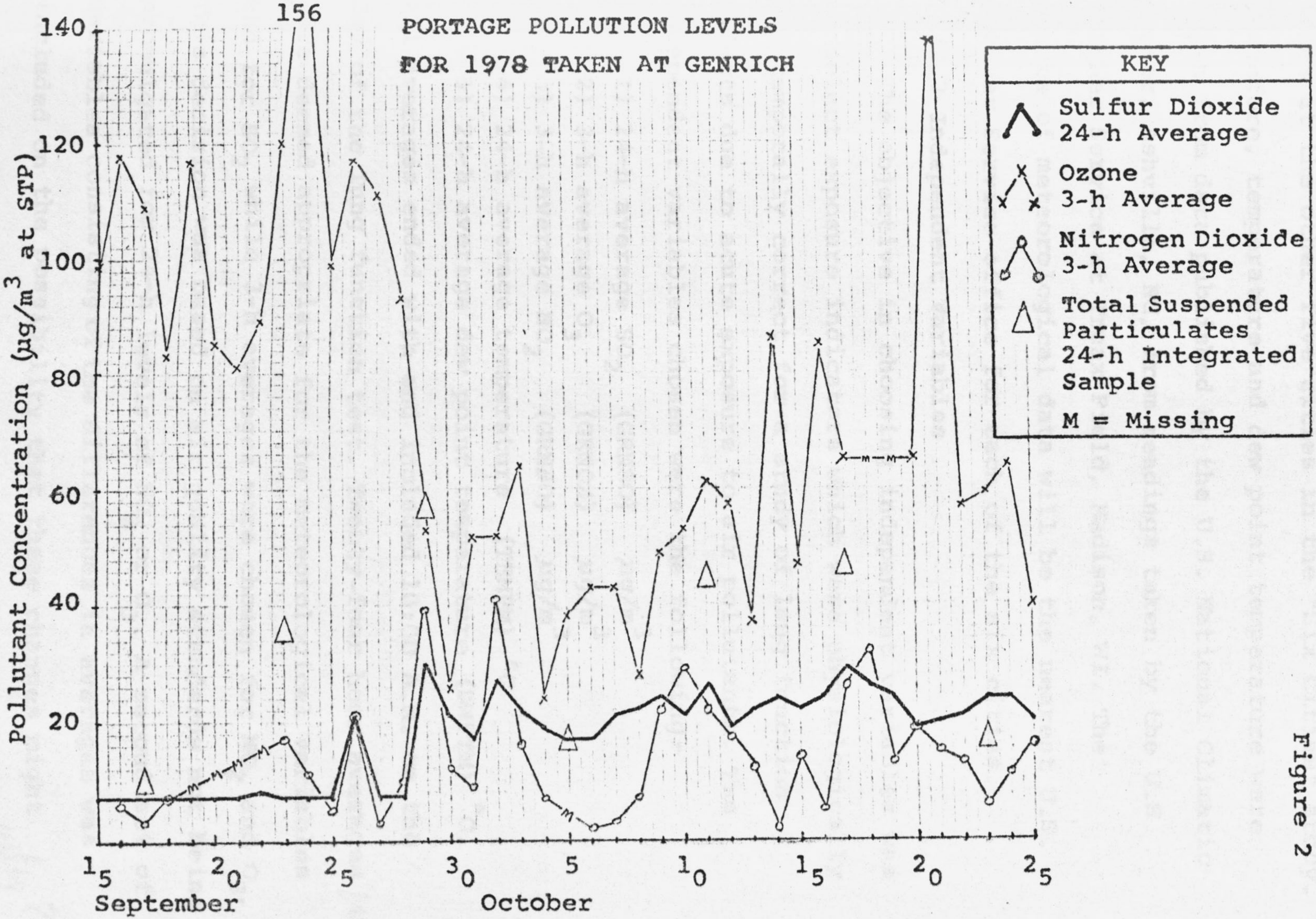

PORTAGE POLLUTION LEVELS
FOR 1978 TAKEN AT GENRICH
156
Pollutant Concentration (μg/m3 at STP)
140
120
100
80
60
40
20
15
20
25
30
5
10
15
20
25
September
October
KEY
Sulfur Dioxide 24-h Average
Ozone 3-h Average
Nitrogen Dioxide 3-h Average
Total Suspended Particulates 24-h Integrated Sample
M = Missing

Figure 2

ities at the other five cities in the "Six Cities" study. Therefore, temperature and dew point temperature were taken from data published by the U.S. National Climatic Center, Ashville, NC, from readings taken by the U.S. Weather Service at Truax Field, Madison, WI. The source of meteorological data will be the nearest U.S. Weather Service office for each of the six cities.

2.7 Independent Variables

The objective in choosing independent variables was to select exposure indicators which were physiologically and temporally correct for a study of lung function effects due to acute exposure to air pollutants. The independent variables chosen were the following:

1) 24-h average SO_2 (GRNSO) $\mu g/m^3$
2) 3-h average O_3 (GRNOZ) $\mu g/m^3$
3) 3-h average NO_2 (GRNNO) $\mu g/m^3$
4) 24-h average temperature (TEMPM) °C
5) 24-h average dew point temperature (DWPTM) °C.

All averages ended with and included 10:00 A.M. on the day of the lung function test. Twenty-four hour averages were deemed appropriate for the meteorological variables and for SO_2 while 3-h averages were chosen for NO_2 and O_3. This decision was based on air quality standards not being promulgated for 24-h levels of NO_2 or O_3. A second set of variables consisting of the differences in averages was included on the possibility that these changes might be more influential in determining lung function than the values themselves. The second data set consists of:

1) change in 24-h average SO_2 from the previous 24-h average (DLTSO)
2) change in 3-h average O_3 from the previous 3-h average (DLTOZ)
3) change in 3-h average NO_2 from the previous 3-h average (DLTNO)
4) change in 24-h average temperature from the previous 24-h average (DLTTM)
5) change in 24-h average dew point temperature from the previous 24-h average (DLTDP).

~~The actual pollutant levels for the 1978 test period, including weekends, are displayed in Figure 2, p. 20.~~ Plots of the two independent data sets described above appear in Appendix B, pp. B1-B5. In these plots the independent variables were matched on the 28 days (COUNT) on which subject 04 was tested and show the approximate matchups which were used in the regressions. Normal plots, histograms and other statistical information concerning these variables appears in Appendix B, pp. B6-B15.

No attempt was made in this research to interpolate missing values among the independent data set. When values were missing the cases were dropped from analysis.

Perhaps the most important statistical consideration with respect to independent variables is their independence. When independent variables are highly correlated with each other a situation of multicolinearity exists. Multicolinear variables are known to produce unstable and sometimes meaningless regression coefficients. In this research the detection of multicolinearity was accomplished by the examination of independent variable correlation matrices and by the evaluation of the relative size of the eigenvalues for the eigenvectors computed from the independent data matrices. When eigenvalues were found that were much smaller than the others and near zero, this was regarded as evidence of multicolinearity. Deletion of variables or eigenvectors was used to control for multicolinear independent variables.

2.8 Regression Procedures

In regressions in which a health indicator is regressed on a group of environmental variables multicolinearity among the independent variables is likely. To select only appropriate predictor variables and to estimate stable regression coefficients, two regression procedures which are especially attuned to dealing with multicolinearity problems were chosen for this analysis. The two procedures were all possible subsets regression (BMDP9R) and regression on principal components (BMDP4R). Comparisons were made between the unbiased results from BMDP9R and the biased results from BMDP4R in order to assess the validity of the findings.

Regression on all possible subsets was selected over stepwise procedures (e.g., BMDP2R) in order to observe the behavior of the regression models with many combinations of the independent variables. BMDP9R selects the "best" subset of the independent variables of each size but also prints out the results of the second best, third best, etc. In this research the criterion by which the "best" subsets were chosen was the maximum adjusted R^2. Adjusted R^2 is equal to $R^2-p(1-R^2)/(N-p')$ where R^2 is the squared multiple correlation, p is the number of independent variables in the model, $p' = p+1$ and N is the number of cases. The "best" subset of any size was also selected by this criterion.

Regression on principal components was selected to both detect and control for multicolinearity. In principal components regression the dependent variable is regressed on linear combinations of the independent variables. These linear combinations or eigenvectors are mutually orthagonal so that all multicolinearity disappears. When all eigenvectors are included in the regression the results are unbiased and the same as those derived from ordinary least squares regression. Bias enters when eigenvectors which explain small amounts of variance in the dependent variables are eliminated. In this research models which adequately described lung function with the fewest number of eigenvectors were sought. After preliminary runs in which all eigenvectors were included, it was decided that eigenvectors with eigenvalues (indications of the amount of variance explained by the eigenvector) less than 1.0 would be eliminated.

Before entering the independent and dependent variables into either regression procedure all variables were standardized using BMDP1S. Although the dependent variables were previously standardized and detrended in their entirety, this second standardization was necessary after case selection was complete. In the future one variable standardization, after case selection, will be used. Standardization of variables was undertaken in order to estimate regression coefficients which were directly comparable for all subjects and for all variables. Even after standardization, however, differences in precision in the regression programs (i.e., BMDP4R single precision and BMDP9R double precision) caused variable means and standard deviations to change slightly from zero and one respectively. For this reason intercepts were not set to zero but were allowed to enter the models. The intercepts were small in all cases and the regression coefficients were close to those expected in a model with zero intercept.

Regression modeling is a multistep procedure with the results of one step indicating to the investigator what should be done in subsequent steps. The first step in this research consisted of running 16 regressions. For each of the four subjects a regression was run for both standardized FEVMN and standardized FVCMN using both BMDP4R and BMDP9R. A combination of both sets of independent variables, totaling 10, was used in each regression.

3. Results

After cases were eliminated because of missing values among the independent variables or because spirometer number 15 was used to measure lung function the following numbers of cases remained: subject 04, n=14; subject 57, n=14; subject 80, n=14; subject 56, n=10. These were regarded as too few cases on which to build meaningful regression equations especially when ten independent variables were under consideration. It was, nevertheless, decided to go ahead with one set of regressions using BMDP9R and BMDP4R to see if any useful information could be gleaned and to identify problems.

Data concerning the overall best subsets of independent variables from BMDP9R are presented in Table 4, p. 27. The best regression equations for subjects 04, 56, and 57 for both STFEV and STFVC were found to be nonsignificant. Of the 25 variables selected for these

TABLE 4 Data on Best Subset Regressions from BMDP9R

STFEV	Subj.	# Var.	Sig. of Regression	Adj. R^2	Significant Variables
	04	3	.4289	.0020	none
	57	4	.1178	.3141	TEMPM
	56	8	.6271	.0982	none
	80	4	.0018	.7489	DLTSO,DWPTM
STFVC					
	04	1	.3760	-.0120	none
	57	1	.1130	.1289	none
	56	8	.3609	.7405	none
	80	7	.0001	.9615	DWPTM,DLTDP, GRNSO,GRNOZ, DLTSO,DLTOZ

models only TEMPM in the regression for STFEV for subject 57 was found to have a regression coefficient significantly different from zero. This was regarded as within the range of random significance. Subject 80 was found to have a significantly predictable STFEV and STFVC by the BMDP9R procedure. Two independent variables, DLTSO and DWPTM, were found to be significant in both regressions.

Data concerning the principal components regressions from BMDP4R are presented in Table 5 below. Only subject 80's

TABLE 5 Data on Principal Components Regression, BMDP4R

STFEV	Subj.	Number of Eigenvectors	Regression F value	R^2	Adj. R^2
	04	3	.18	.0514	-.2332
	57	4	1.45	.3927	.1228
	56	3	.36	.1515	-.2728
	80	4	1.34	.3728	.0940
STFVC					
	04	3	.16	.0452	-.2412
	57	4	.63	.2198	-.1270
	56	3	.75	.2726	-.0911
	80	4	3.71	.6226	.4549

TABLE 6 Subject 80 STFVC

EIGENVECTORS FOR SUBJECT 80 PRINCIPAL COMPONENTS REGRESSION

	1	2	3	4
2 TEMPM	.4125	-.2553	.1820	-.2379
3 DWPTM	.4159	-.1983	.3150	-.1417
4 DLTTM	.2066	-.4850	-.0110	.1002
5 DLTDP	.2453	-.4154	.0934	.1116
6 GRNSO	-.3735	-.2741	.1002	.4372
7 GRNOZ	.2565	-.0427	-.7100	-.3248
8 GRNNO	-.2631	-.4732	.0616	.0219
9 DLTSO	-.3592	-.2851	-.1022	-.3415
10 DLTOZ	.1692	-.1604	-.5445	.5650
11 DLTNO	-.3492	-.2826	-.1845	-.4097
Eigenvalue	3.5932	3.0766	1.2136	1.0910
Cum. Var.	.3593	.6670	.7883	.8974

The remaining eigenvalues and their cumulative proportion of total variance of the independent variables were:

5	.5551	.9530
6	.2517	.9781
7	.1364	.9918
8	.0559	.9973
9	.0217	.9995
10	.0048	1.0000.

The regression equation in terms of the original variables for the four principal component regression was:

$$Y = -.0779 - .0031\text{TEMPM} - .0480\text{DWPTM} - .1416\text{DLTTM} - .1426\text{DLTDP} - .1339\text{GRNSO} + .0738\text{GRNOZ} + .0312\text{GRNNO} + .2246\text{DLTSO} - .4112\text{DLTOZ} + .2664\text{DLTNO}.$$

regression for STFVC was significant at the .05 level. The eigenvectors, eigenvalues and regression coefficients in terms of the original variables for subject 80's regression for STFVC are presented in Table 6 above. The absence of other significant regressions by the principal components method casted doubt on the validity of these results. Without validation from similar findings for other subjects, the regressons found to be significant for subject 80 by the principal components and all possible subsets methods did not inspire confidence in the existence of a relation ship between pollutants and lung function in the samples chosen. This lack of confidence notwithstanding, the independent variables found to be most significant from the regression equation in Table 6, p. 28 and from the best subset for STFVC for subject 80 were DLTSO, DLTOZ, GRNSO, DLTNO and DLTDP. These variables were also consistently among those with the highest correlations with STFVC and STFEV for all subjects. Particular attention will be paid to these variables in future regressions.

Interpretation of eigenvectors is an important aspect of principal components regression. From the eigenvectors displayed in Table 6, p. 28 it can be seen that the eigenvectors, other than number three which might be called the ozone eigenvector, did not offer easy explanation. There was, however, a tendency in all the principal components regressions to find among the top four eignevectors one which was dominated by weather variables, one which was dominated by the effects of sulfur dioxide and nitrogen dioxide and one which was dominated by the effects of ozone.

Eigenvalues as low as .0048 can be seen in Table 6 for eigenvector number 10. This was considered strong evidence of multicolinearity among the independent variables. In these lower eignevectors TEMPM and DWPTM were often found to be dominant. This was not surprising based on the correlation matrix for the variables in the STFVC regression for subject 80 which is presented in Table 7, p.31. The correlation between TEMPM and DWPTM was .9297. It seemed that these two variables should not both be included in the models. Other highly correlated independent variables included DLTTM and DLTDP, DLTSO and DLTNO, and GRNSO and GRNNO. Before final decisions are made on which variables to include and which to eliminate consistent results from analyses of sample sizes larger than 14 must be seen.

No further steps were taken in the regression model building process for two reasons. First, the lack of significant regressions reported above did not point to any significant relationships. Second, this lack of relationships might be entirely understandable in the light of the results presented below.

The correlation matrices of the STFEVs and STFVCs revealed that there was very little correlation between the lung function for the individual subjects.

TABLE 7 Correlation Matrix for Variables in Subject 80 Regression

		STFVC 1	TEMPM 2	DWPTM 3	DLTTM 4	DLTDP 5	GRNSO 6
STFVC	1	1.0000					
TEMPM	2	-.3092	1.0000				
DWPTM	3	-.3240	.9297	1.0000			
DLTTM	4	-.4401	.6329	.4953	1.0000		
DLTDP	5	-.4450	.5496	.5398	.8453	1.0000	
GRNSO	6	.0054	-.4068	-.4085	.1827	.0573	1.0000
GRNOZ	7	-.1838	.3234	.1562	.2469	.2031	-.5268
GRNNO	8	.1871	.0112	-.0634	.4688	.3534	.7425
DLTSO	9	.5507	-.1760	-.2783	.0845	-.1017	.6050
DLTOZ	10	-.4641	.1754	.1564	.3570	.2421	.1041
DLTNO	11	.5468	-.2656	-.3563	.1155	.0139	.4367

		GRNOZ 7	GRNNO 8	DLTSO 9	DLTOZ 10
GRNOZ	7	1.0000			
GRNNO	8	-.2589	1.0000		
DLTSO	9	-.0806	.7107	1.0000	
DLTOZ	10	.3997	.0835	-.1674	1.0000
DLTNO	11	-.0120	.7419	.7843	-.1897

In Table 8, below, only dates on which both subjects were tested were used in computing the correlations. Thus strong correlations should have been present if environmental factors which were the same on each day for each subject, had any consistent influence. The absence of strong correlations was regarded as evidence that the environment played little role in determining lung function during the test period.

TABLE 8 Lung Function Correlations

STFEV Subject	Subject 04	56	57	80	STFEV vs STFVC
04	1.000	-.004	.401	.002	.685
56		1.000	-.367	.235	-.057
57			1.000	.165	.898
80				1.000	.706
STFVC					
04	1.000	-.546	.225	-.394	
56		1.000	-.288	.307	
57			1.000	.209	
80				1.000	

The finding of no relationship between lung function and environmental factors became even more understandable when the lung function coefficients of variation were examined. In Table 9, p. 33, the coefficients of variation which were calculated from raw lung function measurements (i.e., not detrended, not standardized) after case selection, are presented. These variations appeared to be within the range of normal intrapersonal lung function variations for healthy adults.

TABLE 9 Lung Function Coefficients of Variation

Subject	FEVMN	FVCMN
04	.014	.020
56	.026	.024
57	.021	.019
80	.032	.035

4. Discussion

All results from this analysis must be qualified by the unreliability of the lung function data. The detrending of these data to remove "fatigue effects" which was described in section 2.3 and presented graphically in Appendix A, was an unexpected necessity. A "fatigue effect" was not expected based on the findings of Ferris et al. (1978b) and Lawther et al. (1974). It seems possible that either self-administration or administration of the lung function tests by a comrade in a nonclinical setting might have played a role in allowing the significant "fatigue effect" to enter. When autocorrelation due to "fatigue effect" was found in the 1978 Portage data not only the means but the variance, too, became highly suspect. Analysis continued because the purpose of this research was to develop methods, not necessarily to yield results. In a methodological respect the finding of a "fatigue effect" was of importance and it must be watched for carefully in future analysis.

The first objective of this research was to quantify acute lung function effects from air pollutants based on the Portage sample. The null hypothesis for this objective was that all pollutants and meteorological factors have zero effect on lung function. This null hypothesis was not rejected. Despite the significance of subject 80's regression for STFEV and STFVC as displayed in Tables 4 and 5, this finding, based on a sample size of 14, was not accepted as proof of a relationship between lung function and environmental factors in the absence of further validation. Validation in the form of similar results from different subjects was considered crucial to the acceptance of any findings. The use of subjects individually as separate experiments was advocated by Korn and Whittemore (1979) to avoid problems which often occur in panel studies of acute health effects using linear regression. In addition, these individual regressions allow for the cross-validation of regression coefficients for variables of unknown importance. Because of the crudeness of the index of exposure and the effort dependence of the response variable, it was decided in advance that an effect due to a pollutant must be observable in most of the lung function time series in order to be suspected of having a consistent influence.

The finding of no significant influence of lung function due to air pollutants was not surprising. As reported in section 2.5, the pollution levels in the Portage area fall well below current air quality standards. The subjects in this study were healthy adults, not a susceptible population. Under these circumstances a finding of significant lung function impairment would have been alarming. This finding, if it holds up when a more reliable and longer data set is used, would establish Portage as a baseline for lung function effects by which the importance of effects for the same subjects in more polluted environments can be assessed.

The second objective of this research was to select regression procedures and variables to accomplish objective number one. Because there was so little confidence in the validity of the response variable, no steps beyond the preliminary regressions were taken. Therefore the correctness of the model specification remains unknown. It seems appropriate to continue using all possible subsets regression and regression on principal components for reasons stated in section 2.8. Ridge regression is a future possibility if it can offer insights which are not apparent from principal components regression. Biased regression procedures were recommended for analysis in which multicolinear independent variables are a problem by Hocking (1976) and Levenstein et al. (1979). Multicolinearity, based on the results reported in chapter 3, seemed to be an important problem and there is no reason to assume that it will not continue to be troublesome. All possible subsets regression will remain as an unbiased regression procedure with which comparisons can be made.

The selection of independent variables was also an aspect of objective 2. The two meteorological variables were found to be highly correlated. These were 24-h average temperature (TEMPM) and 24-h average dew point temperature (DWPTM). Dew point temperature was selected as a measure of atmospheric moisture instead of relative humidity on the grounds that it is more physically independent of temperature than is relative humidity. Dew point temperature is the temperature at which 100% relative humidity occurs. It is based solely upon the amount of water vapor in the air but is bounded by air temperature in that dew point temperature can be less than or equal to air temperature but can never be greater. Relative humidity is the percentage of water vapor in the air relative to the amount of water vapor that air at that temperature can hold. Therefore cool air which has a low capacity to hold water vapor can have a high relative humidity with little water vapor in the air. If the air temperature rises while the amount of water vapor remains constant, the relative humidity will drop. Since it is hypothesized that water vapor can act in concert with air pollutants to mediate their effects, the more accurate measure of water vapor in the air, dew point temperature, was chosen. The high correlation between temperature and dew point temperature, however, makes the inclusion of both variables in the regressions questionable. This high correlation was also present between the difference in temperature (DLTTM) and the difference in dew point temperature (DLTDP).

During the 1978 Portage test period sulfur dioxide displayed an odd time series of concentrations which is presented in Appendix B, p. B1. Nine out of the first ten readings were 7 $\mu g/m^3$ after which a jump to larger values occurred and continued for the remainder of the test period. Based on SO_2 readings of 7 $\mu g/m^3$ occurring frequently and for prolonged periods both before and after the 1978 test period, it was concluded that these readings were probably not due to instrument or recording error. In section 2.6 the detection limit for SO_2 was stated as less than 10 $\mu g/m^3$ and these readings of 7 $\mu g/m^3$ can probably be regarded as concentrations below the detection limit. One method which was considered for dealing with problems of this type was to assess the effects of SO_2 on lung function both with and without these low readings included in a sort of robust regression. The large number of missing cases which were necessitated for other reasons, however, discouraged further case elimination. In any event, several missing values for NO_2 at the outset of the 1978 test period caused many of these cases to be dropped from analysis before any further specific action could be taken.

All the independent variables are graphed in Appendix B and information concerning their distributions is contained on pp. B6-B15. These data were based on the time series before case elimination due to the use of spirometer 15 which did not affect the environmental variables. The "raw" variables, GRNSO, GRNOZ, TEMPM and DWPTM, all displayed strong trends over the test period; only GRNNO remainded stationary. These trends were not considered desirable attributes for independent variables. This was especially true after the decision was made that trends in the response variable were not real and should be removed. Two variables with trends will always display a strong absolute correlation but this monotonic relationship might often be spurious. Variables which change directions together several times over the course of the experiment are much more convincingly related. For this reason, although there was no indication that the "raw" values were not correct, it was decided to include "deltas" or differences in successive 3-h or 24-h averages. When the "deltas", DLTSO, DLTOZ, DLTNO, DLTTM, and DLTDP, were used as regressors to describe standardized variations in lung function the situation was roughly analagous to the paired-day analysis described by Stebbings (1978). Paired-day analysis involved dividing the data set into pairs of days and defining the variables as the value of the variable on the first day of the pair minus the value on the second day. Stebbings noted that the use of paired-days automatically corrected for the effects of day number. This also occurred when "deltas" were used in this research as is presented in the bottom graphs on pp. B1-B5. The use of "deltas" as regressors and standardized deviations from a zero mean as the response variable allows for the simultaneous direction changes which yield convincing evidence of relationships. These "delta" variables are currently favored as the ones most likely to be used in the acute effects substudy.

Further transformation of the independent variables based on residual analysis after the preliminary regressions was not undertaken because of the small sample sizes (N≤14) and because of the lack of confidence in the response variables as real representations of lung function.

No attempt was made to interpolate missing values among the independent variables in this analysis thus far. Because of the serious problem which resulted from a relative abundance of missing cases, however, attempts might be made to estimate missing values if this problem recurs. The programs BMDPAM and BMDP8D are the procedures which will be used to estimate missing data if it becomes necessary.

Other questions which arise concerning all the independent variables include the appropriateness of the averaging period with respect to the potential physiologic changes and the appropriateness of the fixed ending point for the averages being set at 10:00 AM (9:00 AM for the meteorological variables) for lung function tests ranging from 10:00 AM to 2:00 PM. This fixed ending point could be especially problematic for the 3-h averages used for nitrogen dioxide and ozone.

The third objective in this research was to discover and control significant biasing factors and to identify problems in general. The null hypothesis in this case was that the effect of all potential biasing factors was zero. Only in the case of spirometer number (SPRNO) was the null hypotesis rejected. The finding of consistently low readings for FVC and $FEV_{1.0}$ on spirometer 15 as described in section 2.4 was significant for both the acute effects substudy and for the main "Six Cities Lung Health Study". Harvard investigators had previously found problems in the calibration procedure for their spirometers. The verification of this finding based on a separate data base (i.e., the interviewers rather than the population samples which were analyzed at Harvard) was welcomed. Since 1978 more stringent calibration procedures have been instituted so further spirometer confounding is not expected to affect either the main study or the acute effects substudy. With regard to data collected on malfunctioning spirometers, Harvard investigators are analyzing the error to ascertain whether it is consistent enough to apply a correction factor and thus avoid the loss of a substantial amount of data. Some form of correction will be necessary for this problem if it recurs in the acute effects substudy because case deletion due to missing values has been found to be a major problem.

Other potential biasing factors were not found to be significant. Wind direction (WNDIR), however, will be watched carefully as an indication of the source of pollutants. During the 1978 Portage test period, the wind azimuth, categorized into 22.5 degree groups, varied among seven directions; never blowing out of the northeast. Because of the small sample sizes and the large number of categories (7), the bias testing results were questionable. It is important to know not only the pollutant levels but their sources as well if a broad understanding of the pollution situation for an area is to be attained.

The second major problem found was the serious "fatigue effect" which produced autocorrelation in the response variable. Because this finding has been brought to the attention of the Harvard investigators and subsequently to their field interviewers, it is hoped that it can be avoided by more careful adherence to correct lung function testing protocol by the interviewers. In the event that these decreasing trends due to fatigue persist, four methods seem appropriate for dealing with them. First, as in this present research, the trends can be removed by regressing on test sequence number and the residuals from this regression can be used in further analysis of environmental effects. Second, a variable for test sequence number can be included in the regressions containing the environmental factors. When this was attempted for the 1978 Portage data the sequence number was found to be highly significant while all other variables were found to be nonsignificant. Third, trends in the data may be removed by randomizing the order of the data after matching the response and exposure variables. Subsequent regressions would then be run on the random ordered series. Fourth, paired-day analysis could be adopted in which case effects due to day number would be eliminated. In the unfortunate instance that trends due to effort variablity recur in the interviewer lung function time series, even if detrending can be satisfactorily accomplished, some estimate of the validity of the variations around the detrended series must be made.

Although no significant acute effect due to air pollutants was found on data from the four 1978 Portage subjects, the more important determination was that the data were invalid and as such no believable results were possible. These data, while not reliable for results, were methodologically useful. For example, recommendations offered by Goldstein and Rausch (1978) included ignoring, at first, the independent variables and focusing instead on the dependent variable in an effort to determine whether any significant temporal or geographic deviations from expectations occurred. In other words, find out if there is anything to explain before attempting explanation. In this pilot study of Portage in 1978 the immediate need was to explain the decreasing trends in FVC and $FEV_{1.0}$. Apart from the trend, however, there were variations which might have been environmentally induced. The coefficients of variation for the four subjects' time series after case selection are displayed in Table 9. According to Lawther et al. (1974) normal intrapersonal coefficients of variation for FVC and $FEV_{1.0}$ range from .02 to .04. As can be seen in Table 9, the subjects' lung function did not vary beyond normal limits. This was considered evidence, albeit not proof, that there was nothing to explain. Aside from the variation, means different than expected might also require explanation. The means for FVC and $FEV_{1.0}$ for all subjects were more than one standard deviation away from the means for their height and weight calculated from equations provided by Knudson et al. (1976). In the cases of subjects 04, 56 and 57 the means were above expected while subject 80's mean was below expected. Interpretation of this information was difficult in the light of the apparent meaninglessness of the raw data means. According to Spicer and Kerr (1966), if environmental factors have a consistent effect on respiratory function then effects for individual subjects exposed to the same environment should change as a group. This implies that there should have been strong correlations between the individual subjects' time series. In Table 8 the correlations between lung function time series for the four subjects are presented. The absence of strong correlations was regarded as further evidence that the environment was not a significant influencing factor.

The procedure suggested by Goldstein and Rausch seems an appropriate starting point for the analysis planned for the acute effects substudy. The methodology as it is thus far developed, followed by the programs to be used, is the following:

A) Convert raw lung function data into selected dependent variables by applying BTPS corrections, calculating means of the best three out of five measurements while adhering to the 200ml rule. A FORTRAN program for this purpose has been written.

B) Examine the dependent variables.
   1. Plot lung function measurements and asses the importance of autocorrelation and outliers and remove or control them by the strategies suggested. (BMDP6D)
   2. Test for confounding factors, including seasonality, and control these by a correction factor or by elimination of confounded cases. (BMDP7D)
   3. Re-examine time series with confounding and autocorrelation removed. (BMDP6D)
      a) Determine whether coefficients of variation and and means of lung function series are within normal limits for the whole series and for parts of the time series or for specific geographic locations.
      b) On the basis of "a" above determine whether the lung function time series are valid and whether there are variations which require explanation.
      c) Calculate the correlations between the lung functions of subjects exposed to the same environment and determine whether the environment is a likely factor for explaining variations.
      d) Use the temporal and geographic variations in the lung function series as a guide to suggest environmental variables which might have caused the deviations.

C) Convert environmental data into selected independent variables by calculating 3 or 24-h averages. A FORTRAN program to accomplish this has been written.

D) Match independent variables with dependent variables by eliminating dates on which no lung function tests were performed. (UDH)

E) Examine the independent variables.
   1. Plot the environmental variables and distributional information and look for outliers or transformations which are apparent. (BMDP6D)
   2. Determine the amount of missing data and decide whether interpolation will be necessary. Estimate missing values if needed. (BMDPAM & 8D)
   3. Determine the importance of multicolinearity among the independent variables by analyzing eigenvectors and the independent variable correlation matrix. (BMDP4R)

F) Standardize all dependent and independent variables for cases that remain for regression and create data files for regression input. (BMDP1S)

G) Regress lung function on environmental factors
   1. Run all possible subsets and principal components regressions with various combinations of the independent variables. Pay close attention to envirionmental variables suggested in B3d above. (BMDP9R and BMDP4R)
   2. Analyze models for variable and eigenvector elimination. Analyze residuals for possible outliers and needed transformations among the independent variables.
   3. Compute further regression models with variables transformed and variables and cases deleted as suggested by previous regressions.
   4. Repeat 2 and 3 until models for subjects and cities are correctly specified.

H) Validate results.
   1. Compare results of different regression methods, different cities and different subjects.
   2. Decide whether patterns in the regression coefficients are evident. Determine the amount of lung function variation explanable by environmental factors in each of the six cities and assess the likelihood of the findings being valid.

When this methodology is applied to the interviewer lung function tests in the six cities of the acute effects substudy it would be considered very advantageous, first, if no trends are found in the lung function time series, second, if no spirometers or other confounding factors are found to be significant and third, if there are demonstratable differences in the means and variances of lung function among the six cities. There are reasons to be optimistic about the realization of these three desired conditions.

Trends due to "fatigue effect" and spirometer bias should have been removed due to the notification of the Harvard investigators of their existence and the subsequent corrective actions. With respect to spirometer bias this involved improvements in the calibration procedure. With respect to "fatigue effect" this involved exhortations to the interviewers to adhere carefully to the lung function testing protocol. The interviewer teams whose lung function will be analyzed in the acute effects substudy are not the same as the "children's" interviewer team which was studied in this pilot study. The acute effects substudy will entail the analysis of the lung function tests performed by the "adult" interviewer teams as they make their second visit to each of the six cities. The cities and dates are: Watertown, Massachusetts, fall, 1977; Kinston-Harriman, Tennessee, spring, 1978; St Louis, Missouri, fall, 1978; Steubinville, Ohio, spring, 1979; Portage, Wisconsin, fall, 1979; Topeka, Kansas, spring, 1980. The "adult" teams are both larger than the "children's" teams and remain in each city for a longer period of time. Thus the number of subjects and the length of each time series are increased which improves the chances of valid time series being available for analysis. Hopefully, this will assure valid series even before corrective measures were instituted. In addition, the members of the "adult" interviewer teams are highly motivated individuals who are interested in the acute effects substudy and wish to fully cooperate. This was ascertained upon meetings with the interviewers in Portage in 1979.

Differences in the means and variances of lung function among the six cities are expected based on the wide range of pollutant levels. Portage and Topeka have pollution levels well below National Ambient Air Quality Standards (NAAQS) while St Louis and Steubinville have levels well above. Watertown and Kingston-Harriman are areas of intermediate level pollution. Also, "adult" interviewers spend most of their working days at the local interviewers' office, not at schools as do the "children's" interviewers. This greatly improves the likelihood of accurate exposure estimation, especially when indoor/outdoor pollution monitors were operational at the local interviewers' office. These monitors will provide the badly needed daily readings for particulates, including mass respirable particulates and suspended sulfates. In addition, maximal mid-expiratory flow (MMEF) will be analyzed in the acute effects substudy. MMEF was found to be the most sensitive measure of changes in air pollution by Lawther et al. (1974).

The plan is to apply what was learned in the 1978 Portage pilot study to the improved lung function and environmental data collected for the acute effects substudy. The expectation is that Portage and Topeka testing will display no effects due to pollution while some effects might be present in the more polluted environments of Steubinville and St Louis. As many as ten interviewers have completed lung function testing in some of the cities and three have completed testing in all six cities. Thus an excellant opportunity exists to demonstrate thresholds and possibly a dose-response relationship between lung function variations and acute exposure to air pollutants.

ACKNOWLEDGEMENTS

I would like to thank the members of my committee, Dr. John Rankin of the Department of Preventive Medicine and Dr. John Crowley of the Department of Statistics of the University of Wisconsin, for reviewing this work. Special gratitude is extended to Dr. Marty Kanarek of the Department of Preventive Medicine for serving as thesis chairman. I would also like to thank the members of the "Six cities Lung Health Study" at the Harvard University School of Public Health for their assistance. I would especially like to acknowledge the help provided by Dr. Doug Dockery in coordinating my project within the main study. The Water Resources Center of the University of Wisconsin, headed by Dr. Gordon Chesters, provided essential data and arranged funding for this research through a grant from the Wisconsin Power and Light Company. I wish to thank Gordy Thomas of the Department of Preventive Medicine for aid in computer programming and Sheila Timme of the Water Resources Center for general assistance. Last, but certainly not least, I would like to thank the Harvard field interviewers for completing many tedious spirometric tests.

REFERENCES

BMDP Biomedical Computer Programs, P Series, Health Science Computing Facility, Department of Biomathematics, The University of California, Los Angeles, University of California Press, 1979.

Chatterjee, S., and B. Price, Regression Analysis by Example, John Wiley & Sons, New York, 1977.

Cotes, J.E., Lung Function, Blackwell Scientific Publications, Oxford, 1975.

Draper, N, and H. Smith, Applied Regression Analysis, John Wiley & Sons, New York, 1966.

Ferris, Benjamin G., Jr., Health Effects of Exposure to Low Levels of Regulated Air Pollutants, A Critical Review, Air Pollution Control Association Journal, 28: 482-497, 1978.

Ferris, Jr., B.G., F.E. Speizer, J.D. Spengler, D. Dockery, Y.M.M. Bishop, M. Wolfson and C. Humble, "Effects of Sulfur Oxides and Respirable Particulates on Human Health, 1. Methodology and Demography of Populations in Study", Harvard School of Public Health, 1978a.

Ferris, Jr., B.G., F.E. Speizer, Y. Bishop, G. Prang, and J. Weener, Spirometry for an Epidemiologic Study: Deriving Optimum Summary Statistics for Each Subject, Bull. Europ. Physiopath. Resp., 14: 145-166, 1978b.

Ferris, B.G., Jr., Principal Investigator, Epidemiology Standardization Project, American Review of Resp. Dis., 118(6):55-88, 1978c.

Goldstein, I.F., and L.E. Rausch, Time Series Analysis of Morbidity Data for Assessment of Acute Environ mental Health Effects, Environ. Res., 17:266-275, 1978.

Hocking, R.R., The Analysis and Selection of Variables in Linear Regression, Biometrics, 32:1-49, 1976.

Hosein, H.R., C.A. Mitchell and A. Bouhuys, Daily Variation in Air Quality, Arch. Environ. Health, January/February 1977, pp. 14-21.

Kerr, H.D., T.J. Kulle, M.L. McIlhany and P. Swidersky, Effects of Ozone on Pulmonary Function in Normal Subjects, Am. Rev. Resp. Dis., 111:763-773, 1975.

Knudson, R.J., R.C. Slatin, M.D. Lebowitz and B. Burrows, The Maximal Expiratory Flow-Volume Curve, Am. Rev. Resp. Dis., 113:587-600, 1976.

Korn, E.L., and A.S. Whittemore, Methods for Analyzing Panel Studies of Acute Health Effects of Air Pollution, Biometrics, 35:795-802, 1979.

Lawther, P.J., A.G.F. Brooks, P.W. Lord and R.E. Waller, Day to Day Changes in Ventilatory Function in Relation to the Environment, Part 1. Spirometric Values, Environ. Res., 7:27-40, 1974.

Lawther, P.J., A.J. MacFarlane, R.E. Waller and A.G.F. Brooks, Pulmonary Function and Sulfur Dioxide, Some Preliminary Results, Environ. Res., 10:355-367, 1975.

Lebowitz, M.D., E.J. Cassel, J.D. McCarroll, Health and the Environment. XV. Acute Respiratory Episodes as Reactions by Sensitive Individuals to Air Pollution and Weather, Environ. Res., 5:135-141, 1972.

Levenstein, M.J., Y.M.M. Bishop, B.G. Ferris, Jr., F.E. Speizer, Six Cities Study: Standardization of Pulmonary Function Measurements, In: Energy and Health Science, Sims Conference Series, Number 6, Philadelphia: SIAM, 1979.

Linn, W.S., R.D. Buckley, C.E. Spier, R.L. Blessey, M.P. Jones, D.A. Fischer and J.D. Hackney, Health Effects of Ozone Exposure in Asthmatics, Am. Rev. Resp. Dis., 117: 835-843, 1978.

Munn, R.E., Biometeorological Methods, Academic Press, New York, 1970.

National Academy of Sciencs, Ozone and Other Photochemical Oxidants, Medical and Bidogic Effects of Environmental Pollutants, N.A.S., Washington, D.C., 1977a.

National Academy of Sciences, Nitrogen Oxides, Medical and Biologic Effects of Environmental Pollutants, N.A.S., Washington, D.C., 1977b.

National Academy of Sciences, Sulfur Oxides, Medical and Biologic Effects of Environmental Pollutants, N.A.S., Washington, D.C., 1978.

National Academy of Sciences, Airborne Particles, Medical and Biologic Effects of Environmental Pollutants, N.A.S., University Park Press, Baltimore, 1979.

Speizer, F.E., Y. Bishop and B.G. Ferris, Jr., An Epidemiologic Approach to the Study of Health Effects of Air Pollution, In: Proceedings of the 4th Symposium on Statistics and the Environment, March 3 5, 1976, Washington, D.C.: American Statistical Association, pp. 56-68, 1977.

Spicer, W.S., Jr., and D.H. Kerr, Variation in Respiratory Function, Arch. Environ. Health, 12:217-226, 1966.

Stebbings, J.H., Jr., Panel Studies of Acute Health Effects of Air Pollution. II. A Methodologic Study of Linear Regression Analysis of Asthma Panel Data, Environ. Res., 17:10-32, 1978.

Tager, I., F.E. Speizer, B. Rosner and G. Prang, A Comparison between the Three Largest and Three Last of Five Forced Expiratory Maneuvers in a Population Study, Am. Rev. Resp. Dis., 114:1201-1203, 1976.

APPENDIX A

Data on the

Dependent Variables

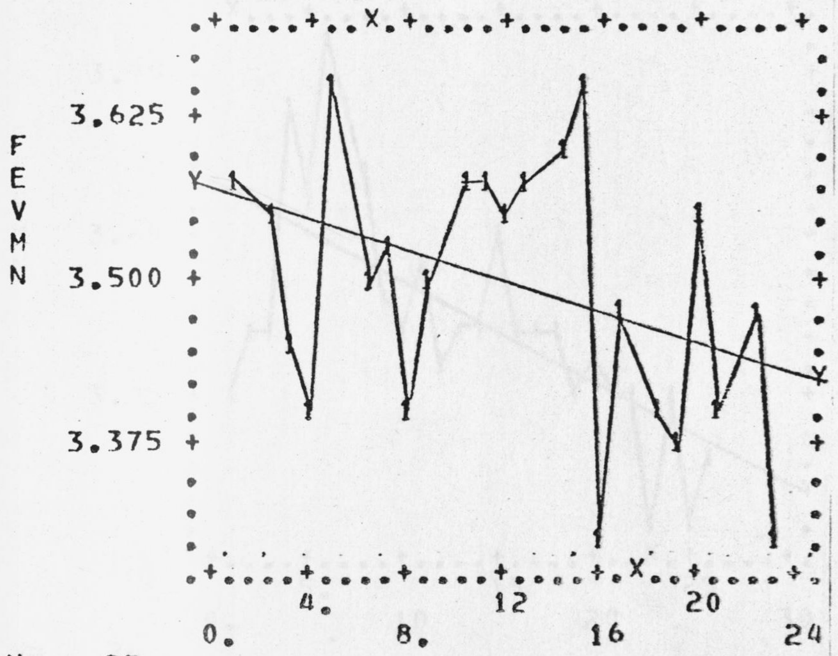
KEY

N = the number of cases
COR = correlation between lung function on the y axis and the sequence number of the test on the x axis.
COUNT = the sequence number of the test
FVCMN = the mean of the best three out of five FVC readings for each day's test
FEVMN = the mean of the best three out of five FEV readings for each day's test
STFVC = standardized FVCMN after the trend is removed
STFEV = standardized FEVMN after the trend is removed

"Original" indicates graphs in which lung function is expressed in units of liters.

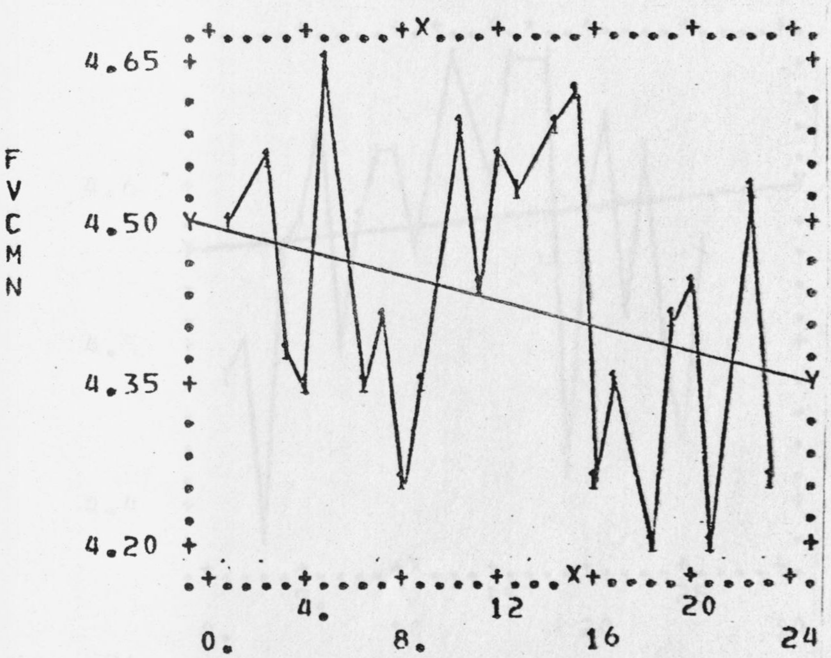
"Standardized" indicates graphs in which lung function has been detrended and expressed in terms of deviations from a regression line divided by the standard error of the residual. Units are residual standard errors.

Subject 56 Original



N= 23
COR=-.396

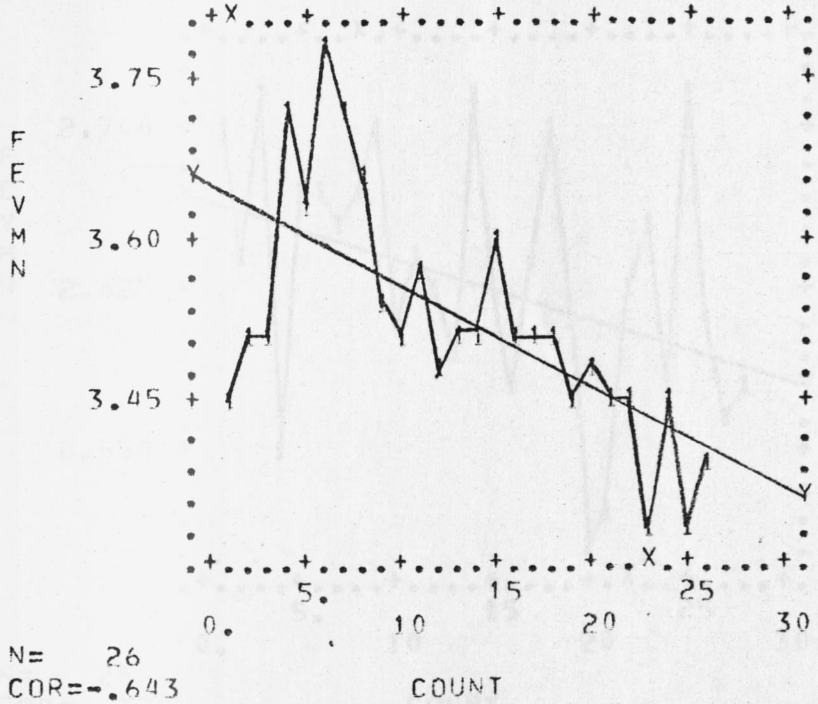
COUNT



N= 23
COR=-.269

COUNT

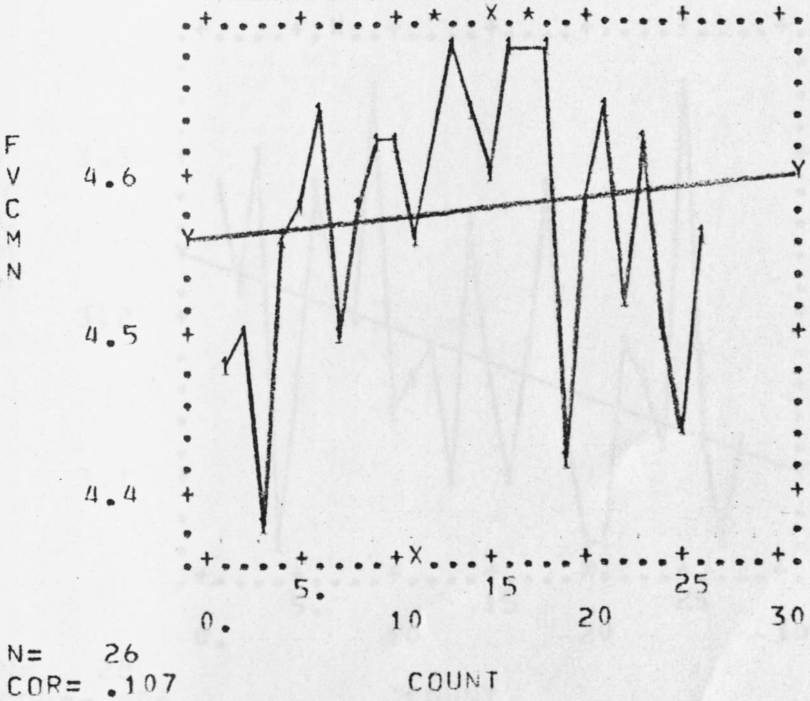
Subject 57 **Original**



N= 26
COR=-.643

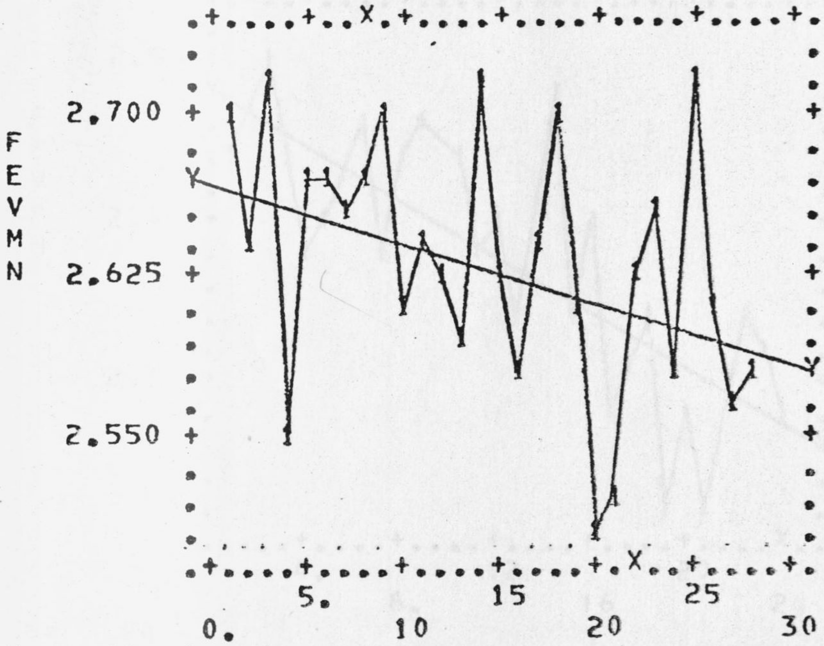
FEV/FVC decreasing

? explanation



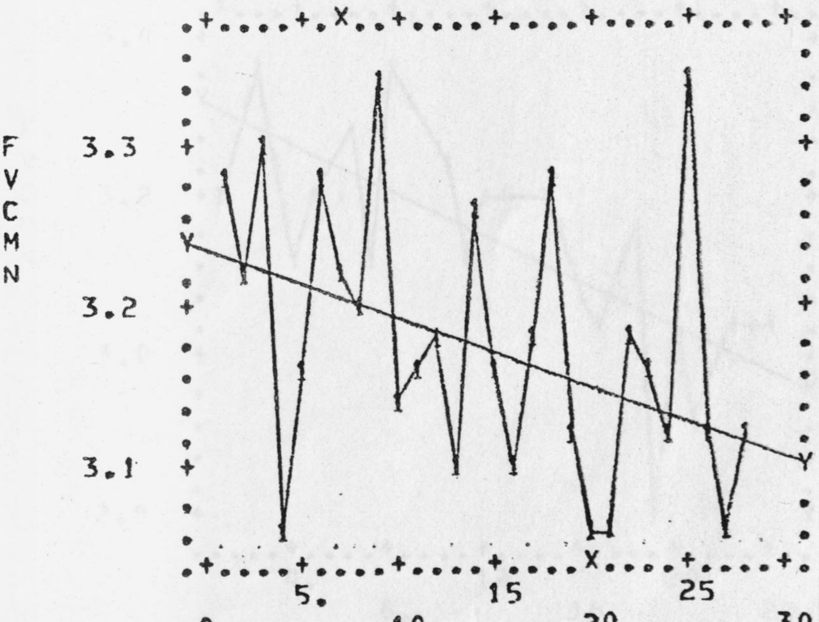
N= 26
COR= .107

Subject 04 Original



N= 28
COR=-.390

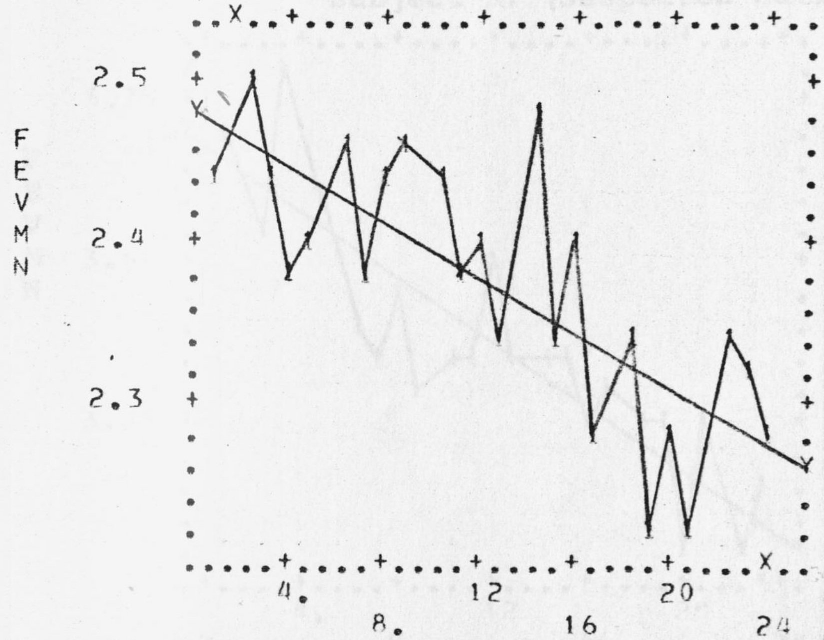
COUNT



N= 28
COR=-.406

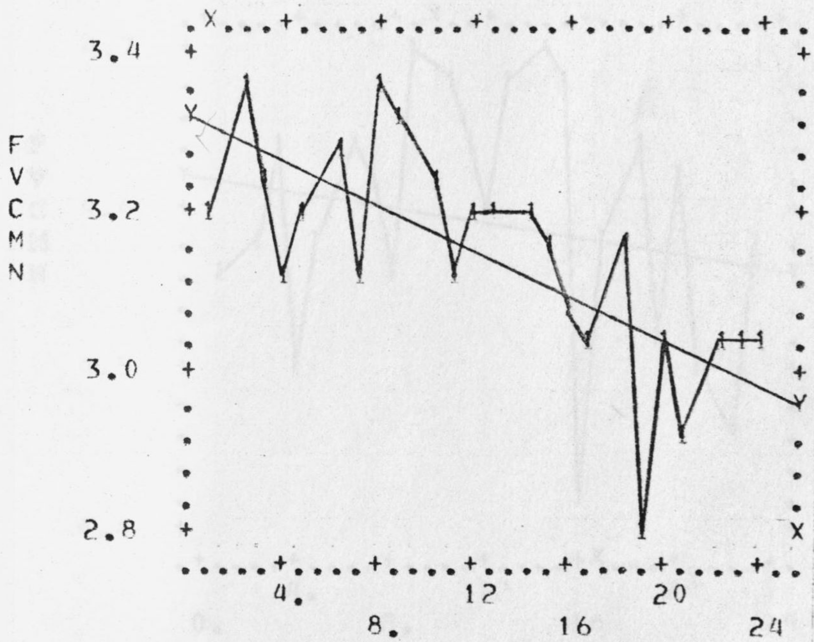
COUNT

Subject 80 Original



N= 24
COR=-.762

COUNT

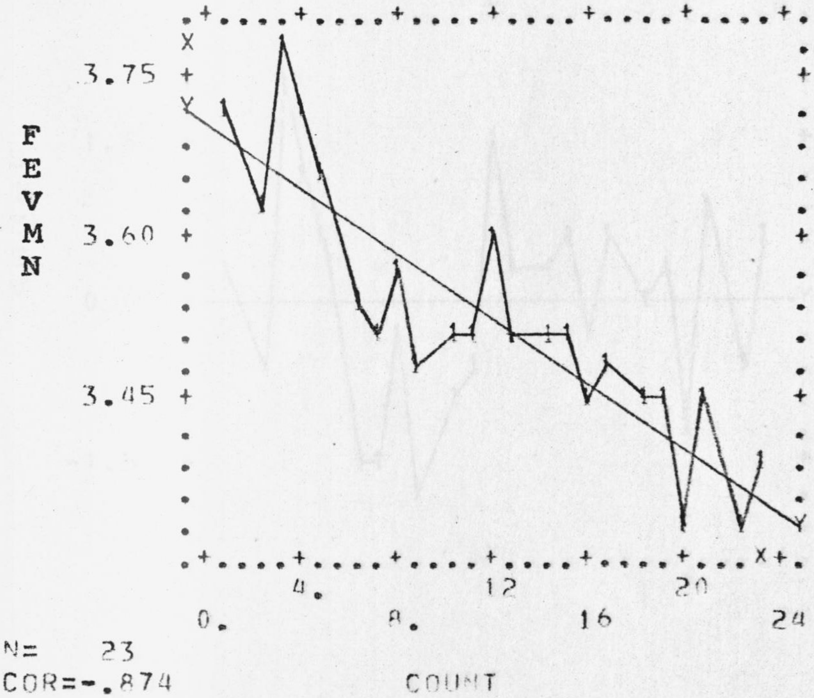


N= 24
COR=-.708

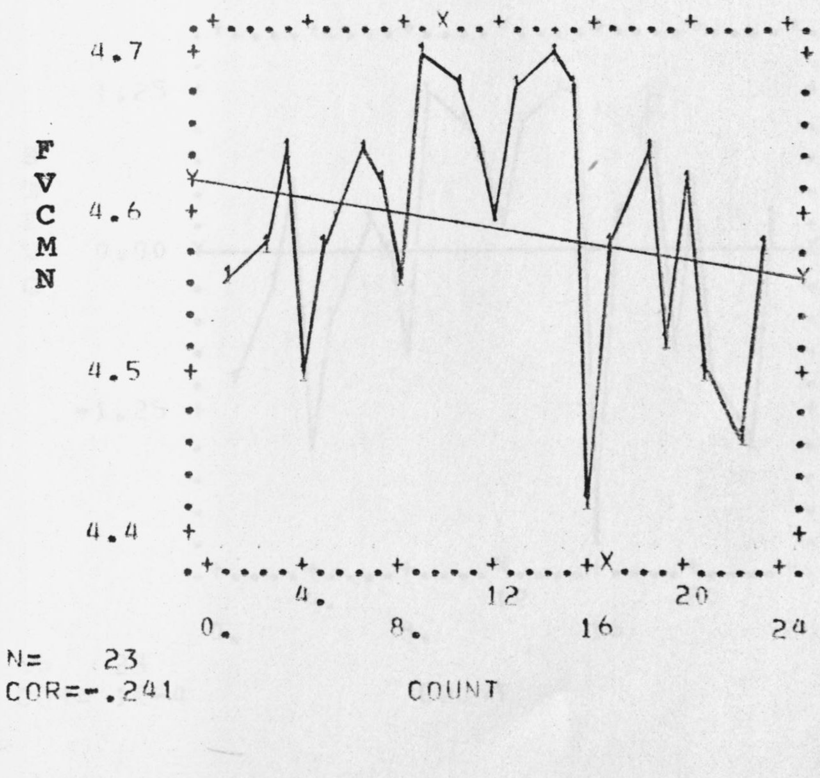
COUNT

.3 in 2 quarters
1.2 in one year
normal 30ml/yr

Subject 57 (Suspected outliers removed)



N= 23
COR=-.874

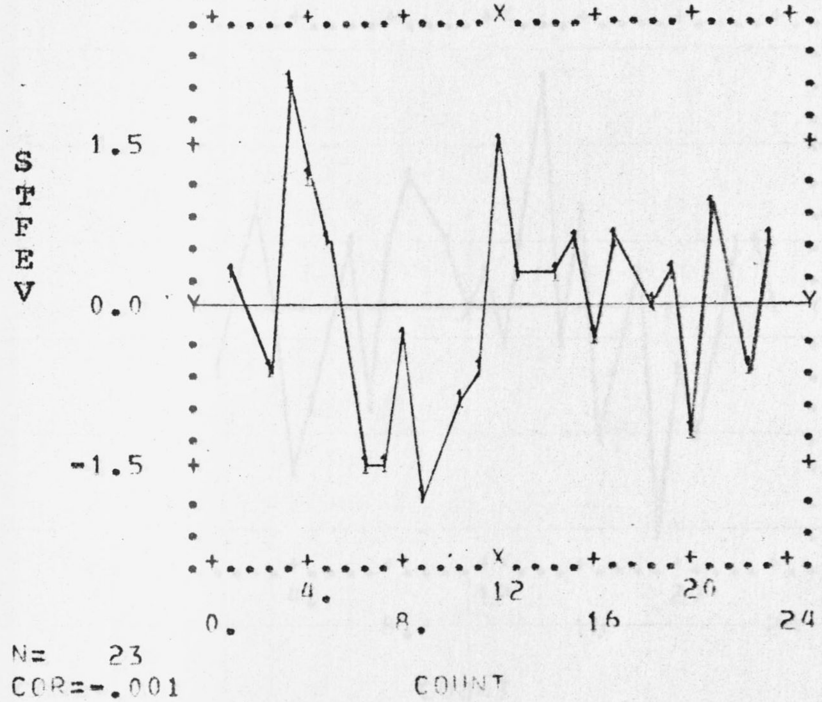


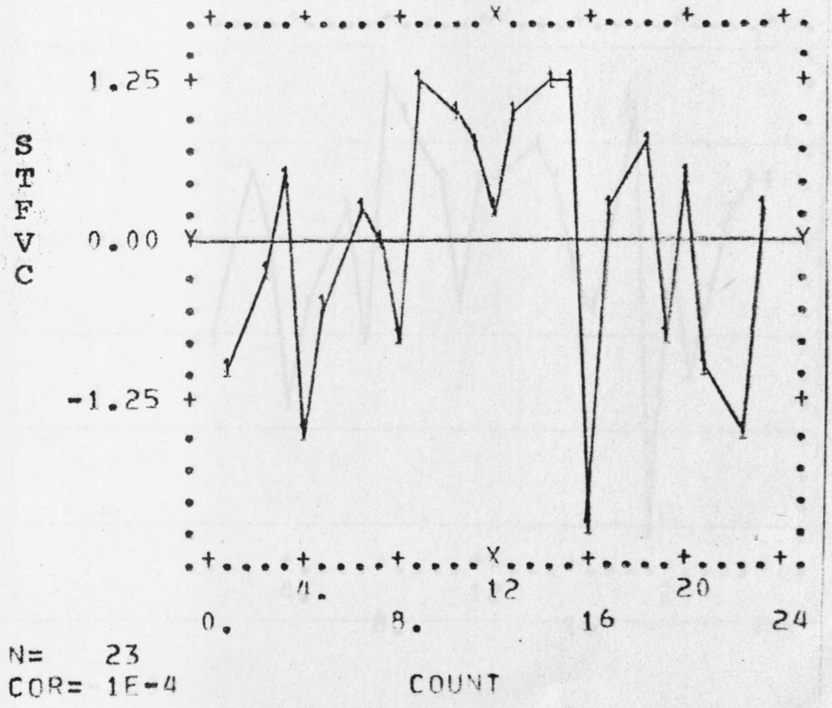
N= 23
COR=-.241

FEV/FVC decreasing

Subject 57

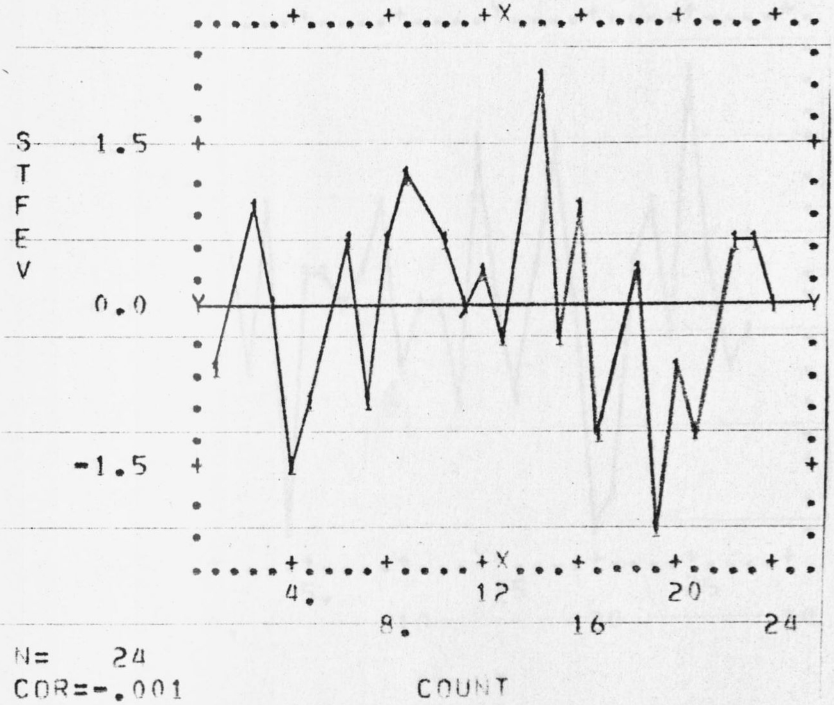
Standardized
(Outliers
removed)

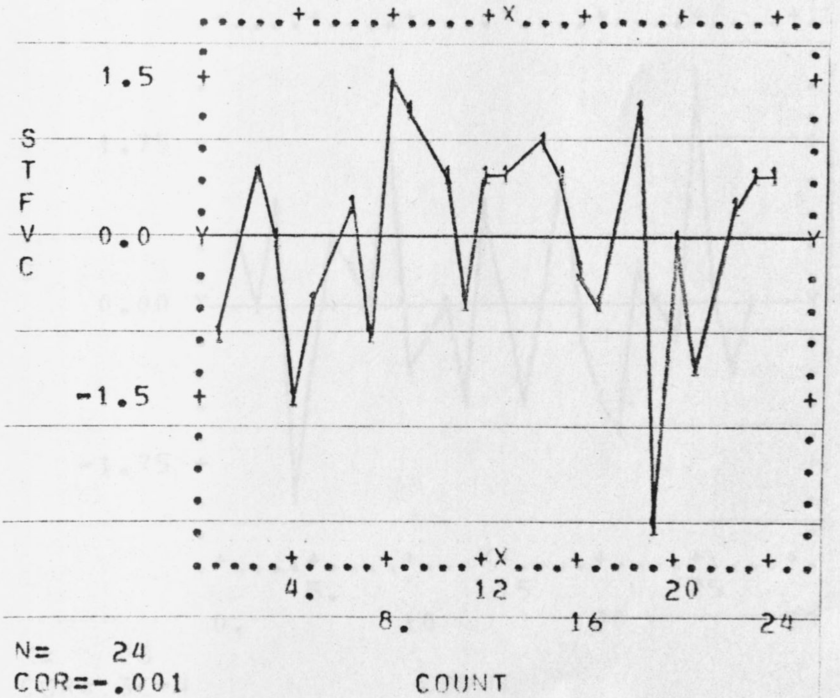




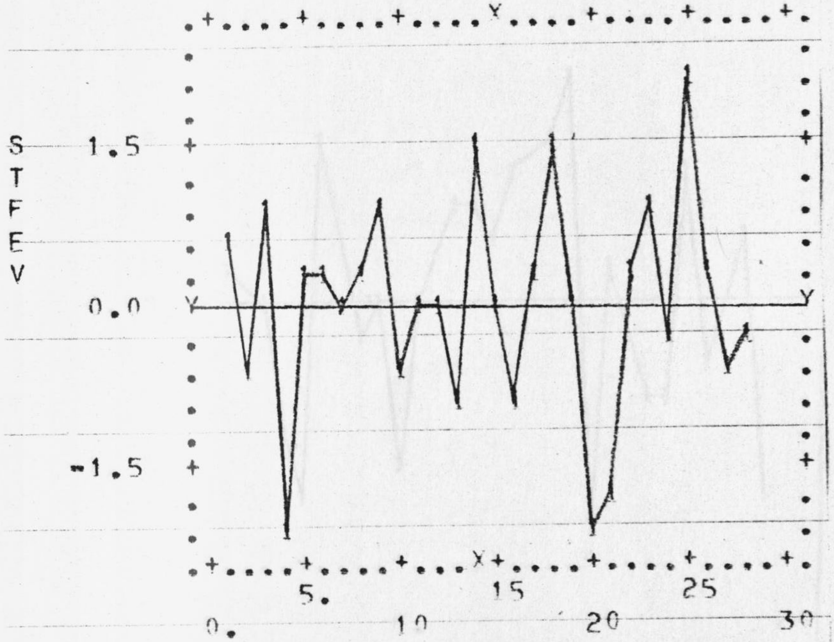
Subject 80

Standardized



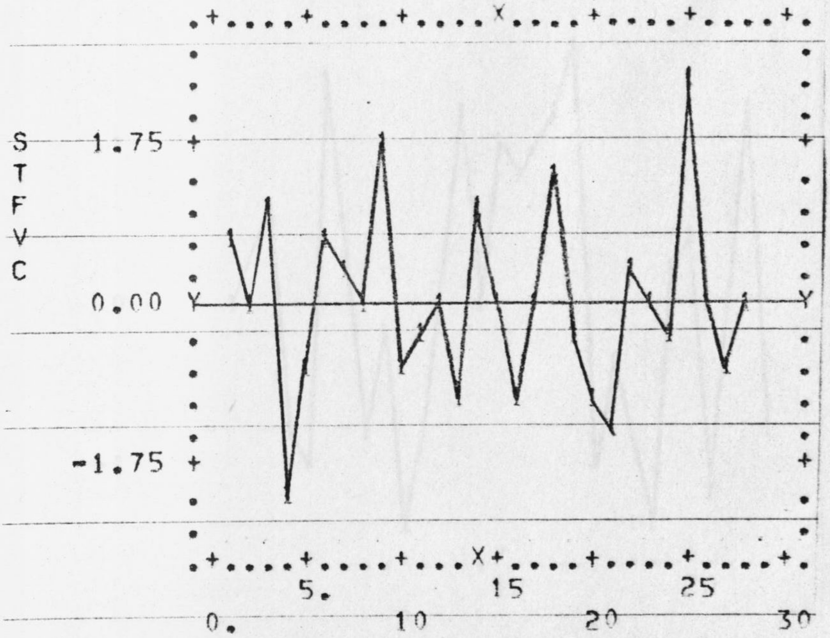


Subject 04 Standardized



N= 28
COR= .001

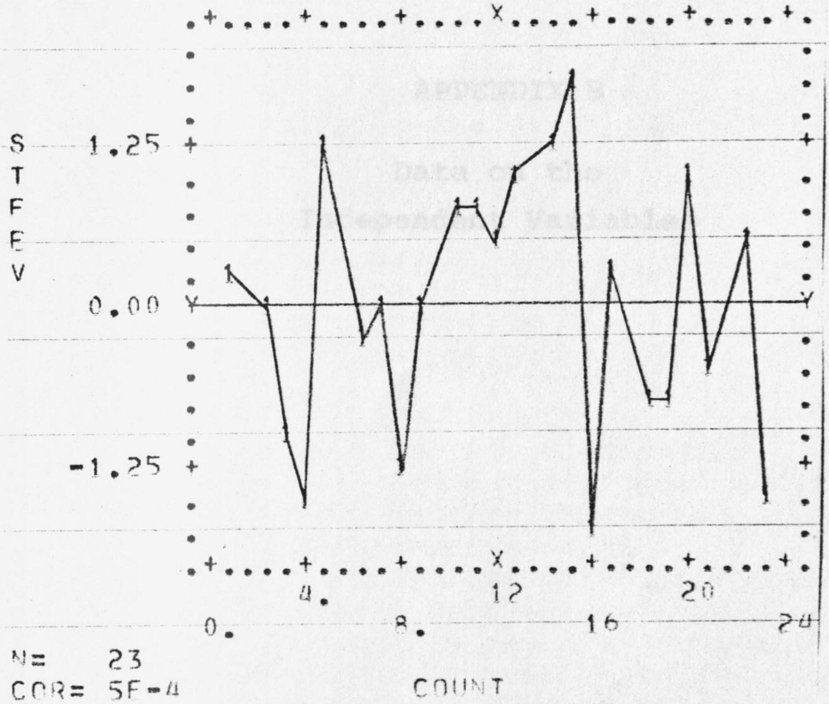
COUNT

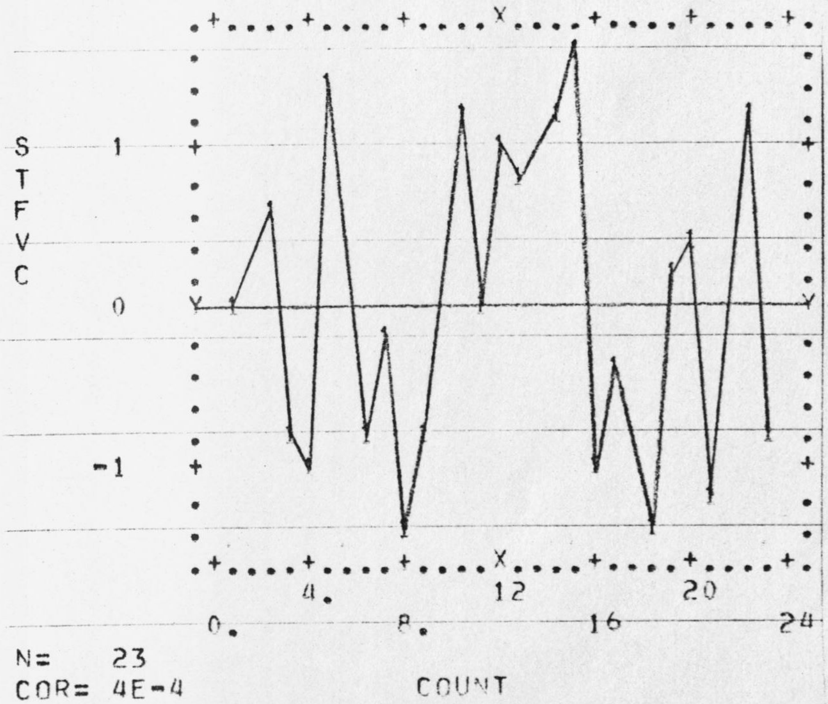


N= 28
COR= 3E-4

COUNT

Subject 56 Standardized





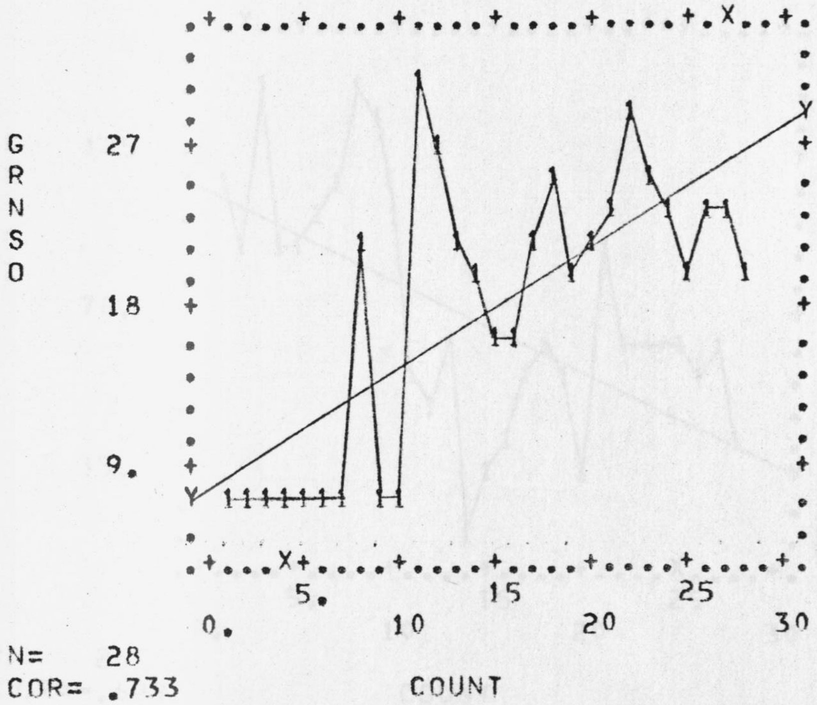
APPENDIX B

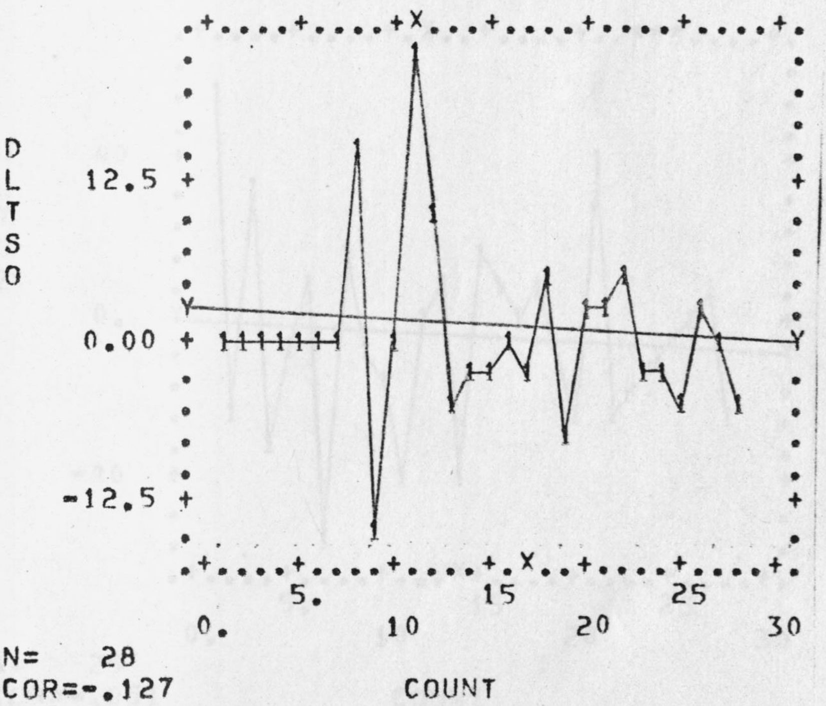
Data on the Independent Variables

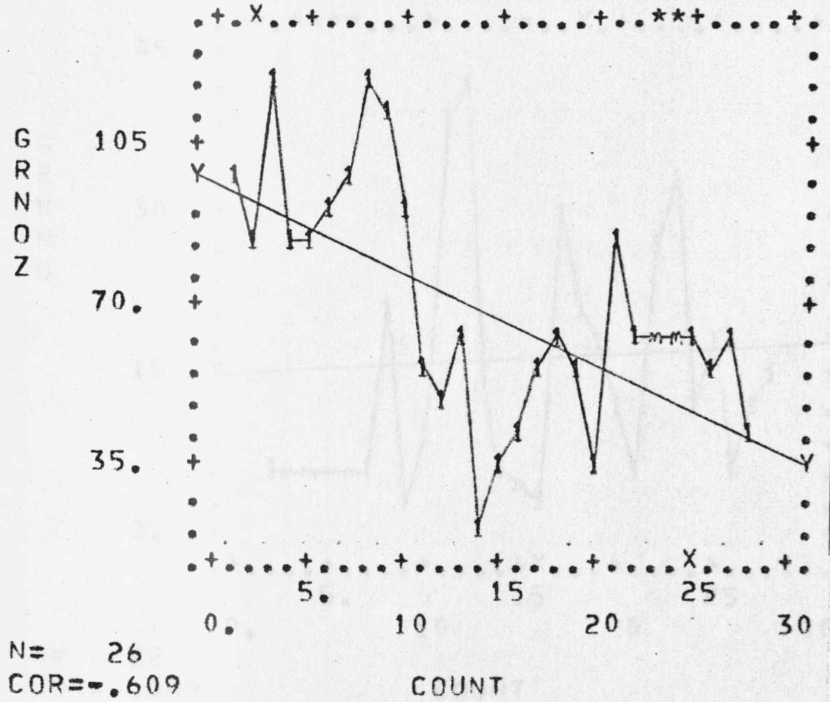
KEY

N	=	the number of cases on which the statistics were computed
X	=	mean
s	=	standard deviation
CV	=	coefficient of variation
skew.	=	skewness
kurt.	=	kurtosis
min.	=	minimum value
max.	=	maximum value

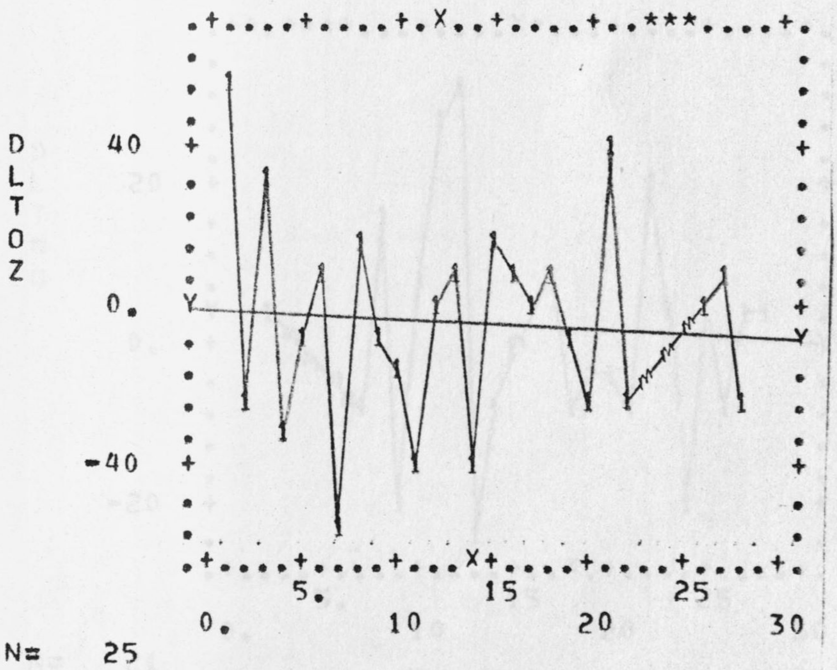
TEMPM	=	24-h average temperature
DWPTM	=	24-h average dew point temperature
GRNSO	=	24-h average sulfur dioxide
GRNOZ	=	3-h average ozone
GRNNO	=	3-h average nitrogen dioxide
DLTTM	=	change in average temperature
DLTDP	=	change in average dew point temperature
DLTSO	=	change in average sulfur dioxide
DLTOZ	=	change in average ozone
DLTNO	=	change in average nitrogen dioxide


GRNSO
27
18
9.
0.
5.
10
15
20
25
30
N= 28
COR= .733
COUNT

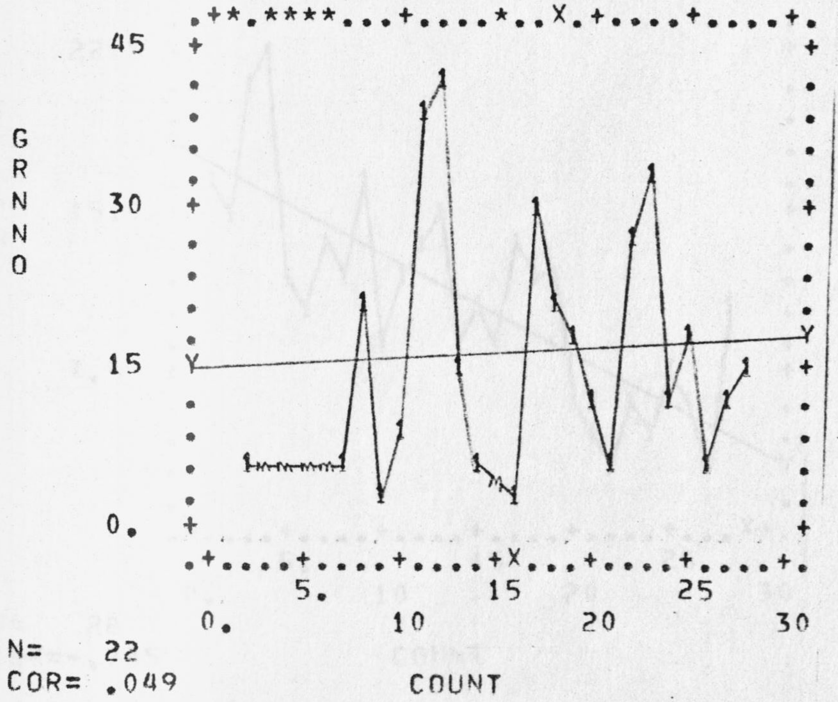

DLTSO
12.5
0.00
-12.5
0.
5.
10
15
20
25
30
N= 28
COR=-.127
COUNT

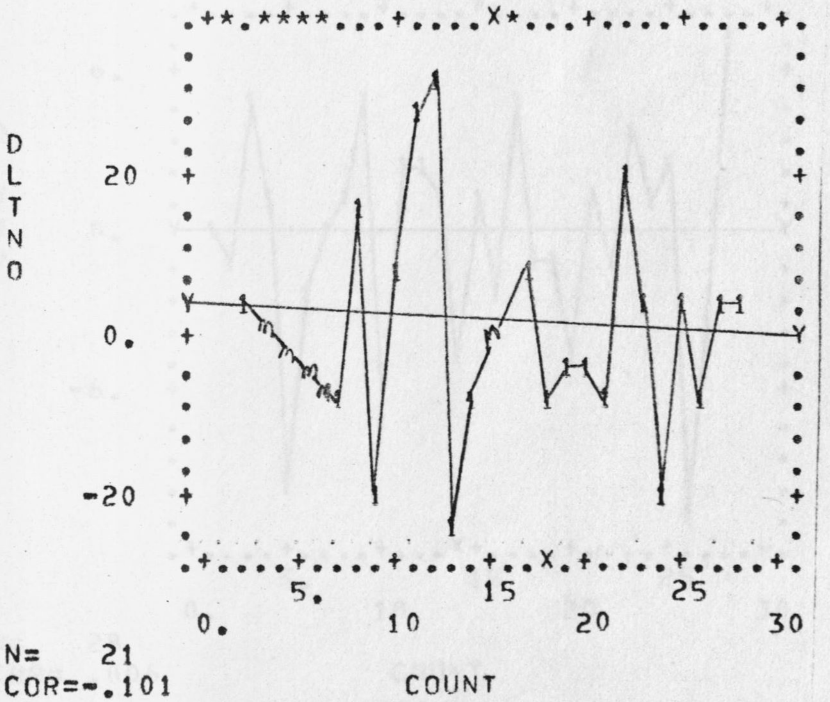


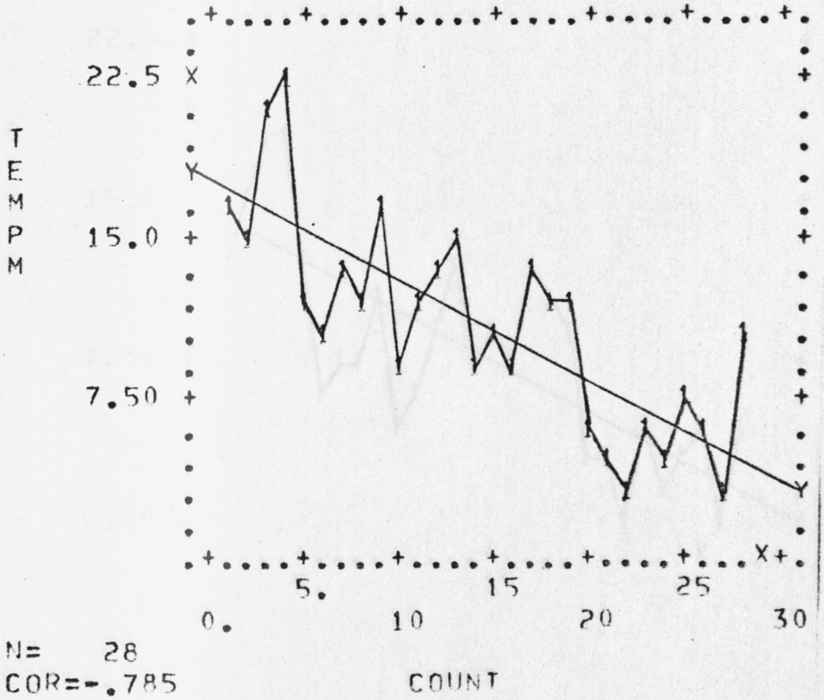
N= 26
COR=-.609

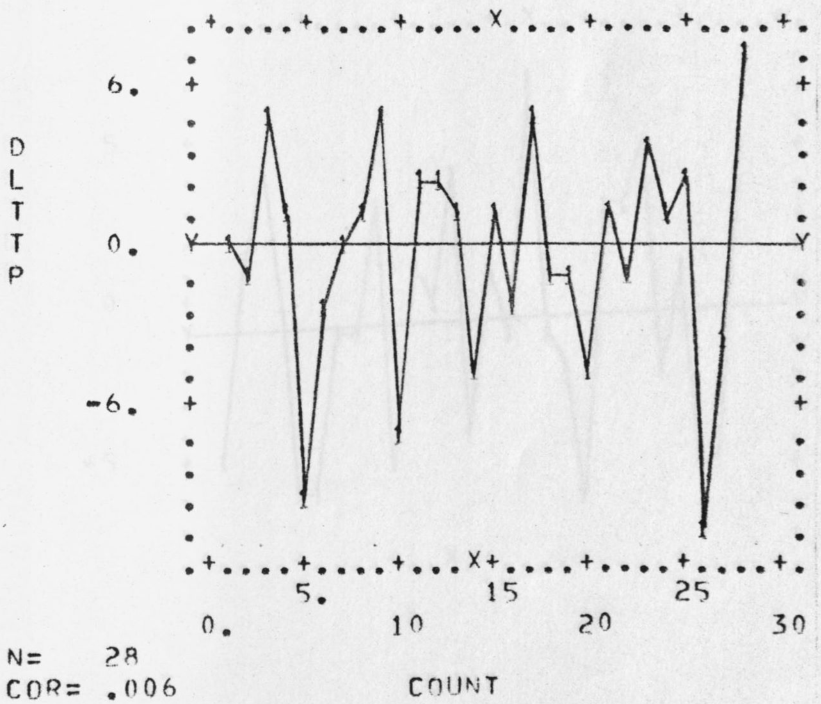


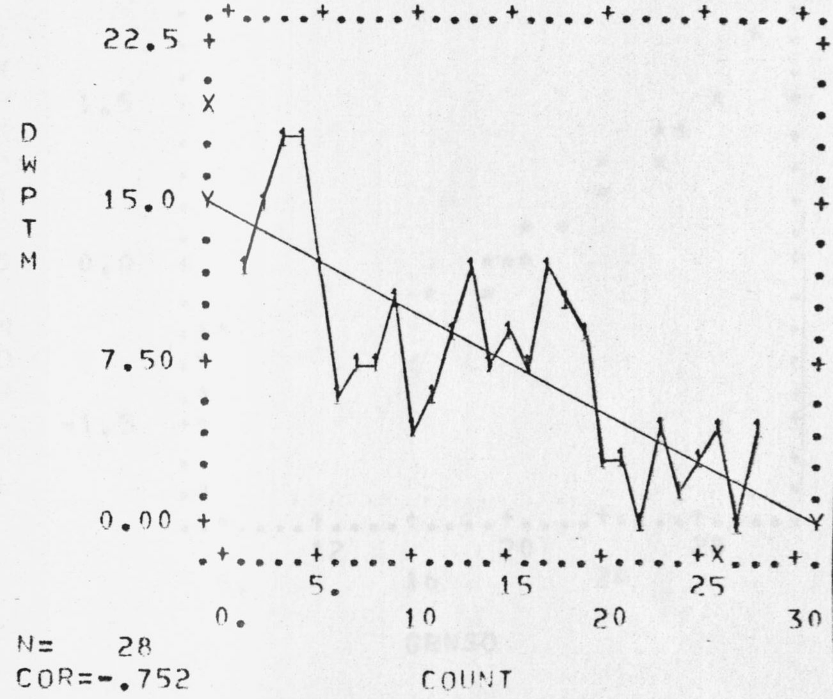
N= 25
COR=-.051

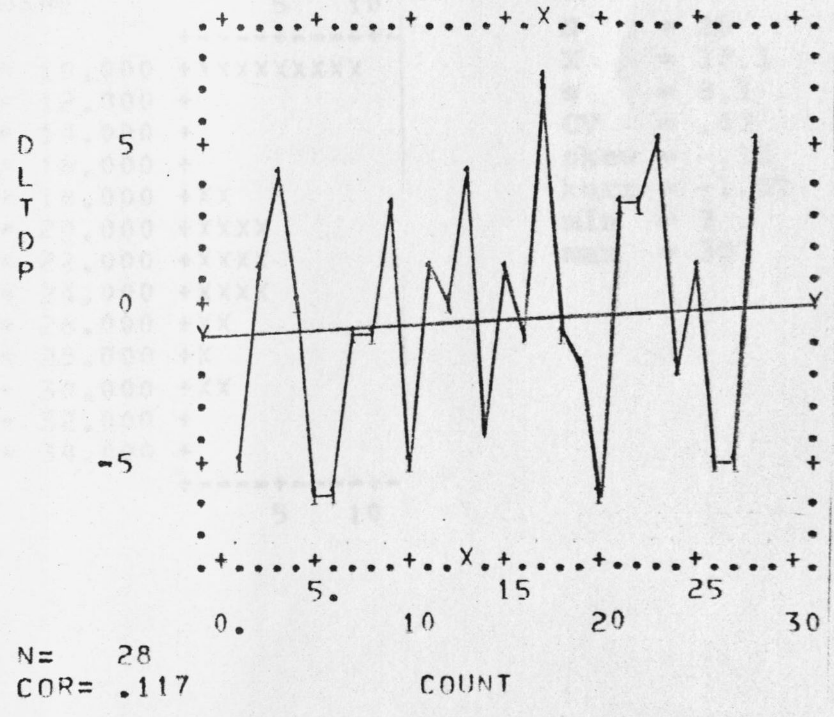

45
30
15
0.
G
R
N
N
O
5.
15
25
0.
10
20
30
N= 22
COR= .049
COUNT


20
0.
-20
D
L
T
N
O
5.
15
25
0.
10
20
30
N= 21
COR=-.101
COUNT

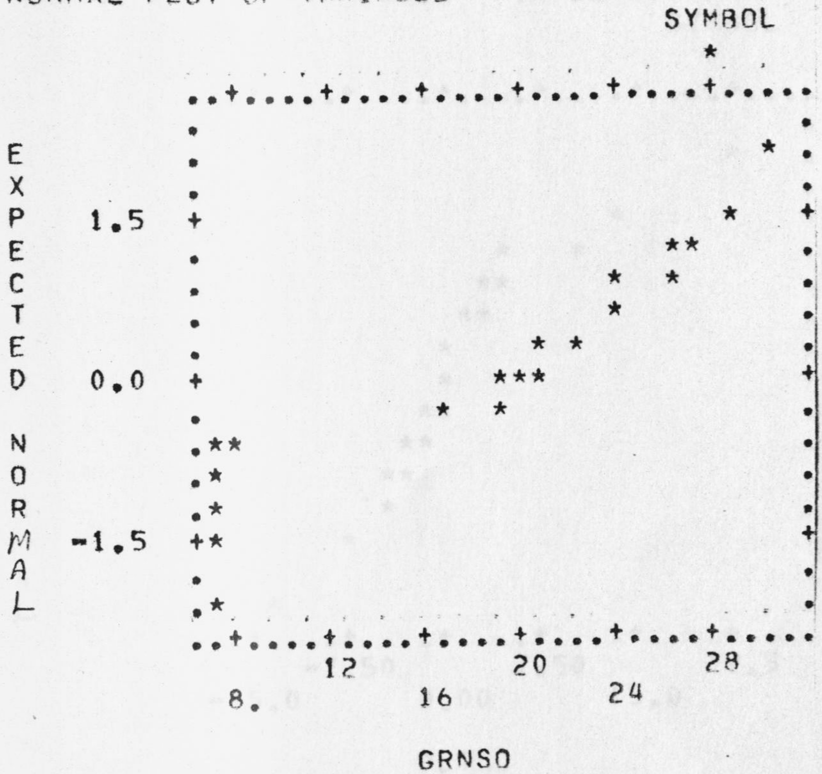

22.5
15.0
7.50
TEMPM
0.
5.
10
15
20
25
30
N= 28
COR=-.785
COUNT


6.
0.
-6.
DLTTP
0.
5.
10
15
20
25
30
N= 28
COR= .006
COUNT


22.5
15.0
7.50
0.00
D
W
P
T
M
5.
15
25
0.
10
20
30
N= 28
COR=-.752
COUNT


5
0
-5
D
L
T
D
P
5.
15
25
0.
10
20
30
N= 28
COR= .117
COUNT

NORMAL PLOT OF VARIABLE 8 GRNSO




```
INTERVAL
NAME              5    10
                +----+----+-
* 10.000  +XXXXXXXXX
* 12.000  +
* 14.000  +
* 16.000  +
* 18.000  +XX
* 20.000  +XXXX
* 22.000  +XXXX
* 24.000  +XXXX
* 26.000  +XX
* 28.000  +X
* 30.000  +XX
* 32.000  +
* 34.000  +
                +----+----+-
                  5    10
```

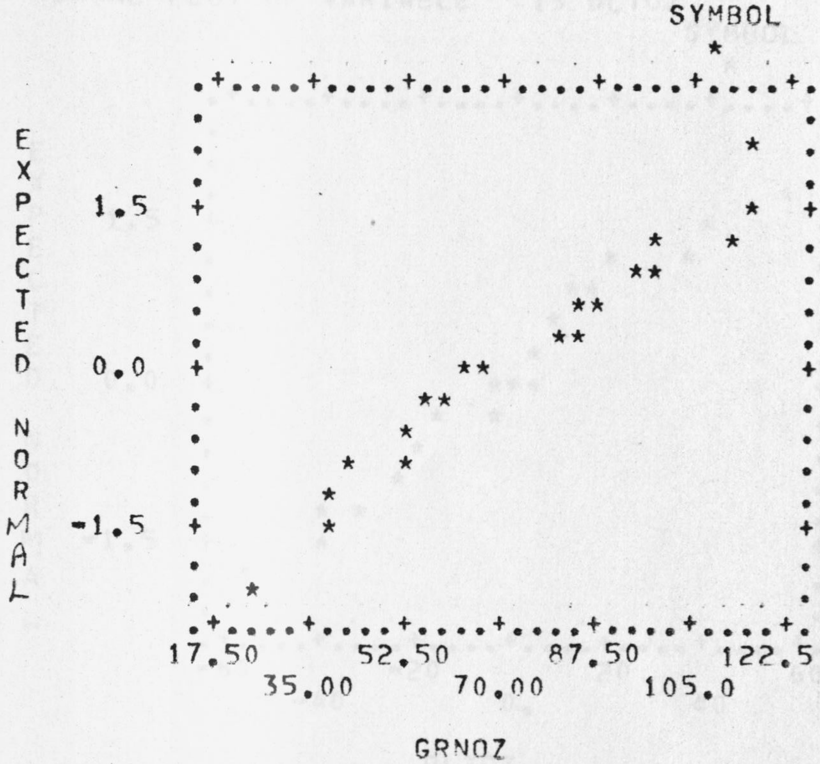
N = 26
X = 17.1
s = 8.1
CV = .47
skew = -.16
kurt = -1.52
min = 7
max = 30

```
NORMAL PLOT OF VARIABLE   14 DLTSO
                                          SYMBOL
                                            *

          ...+....+....+....+....+....+....
          .                               .
E         .                           *   .
X         .                               .
P   1.5   +                       *       +
E         .                 *   *         .
C         .                **             .
T         .               **              .
E         .               *               .
D   0.0   +               *               +
          .              **               .
N         .             **                .
O         .            **                 .
R         .            *                  .
M  -1.5   +         *                     +
A         .                               .
L         .   *                           .
          ...+....+....+....+....+....+....
               -7.50     7.50      22.5
          -15.0     0.00      15.0

                     DLTSO
```

```
INTERVAL
NAME              5   10   15
            +----+----+----+-
*-10.500    +X
*-7.0000    +X
*-3.5000    +XXX
* 0.0000    +XXXXXXXXXXXXXX
* 3.5000    +XXXX
* 7.0000    +XX
* 10.500    +X
* 14.000    +X
* 17.500    +
* 21.000    +
* 24.500    +X
* 28.000    +
* 31.500    +
            +----+----+----+-
                  5   10   15
```

N = 25
X = 1.1
s = 7.0
CV =
skew = 1.04
kurt = 2.39
min = −14
max = 23

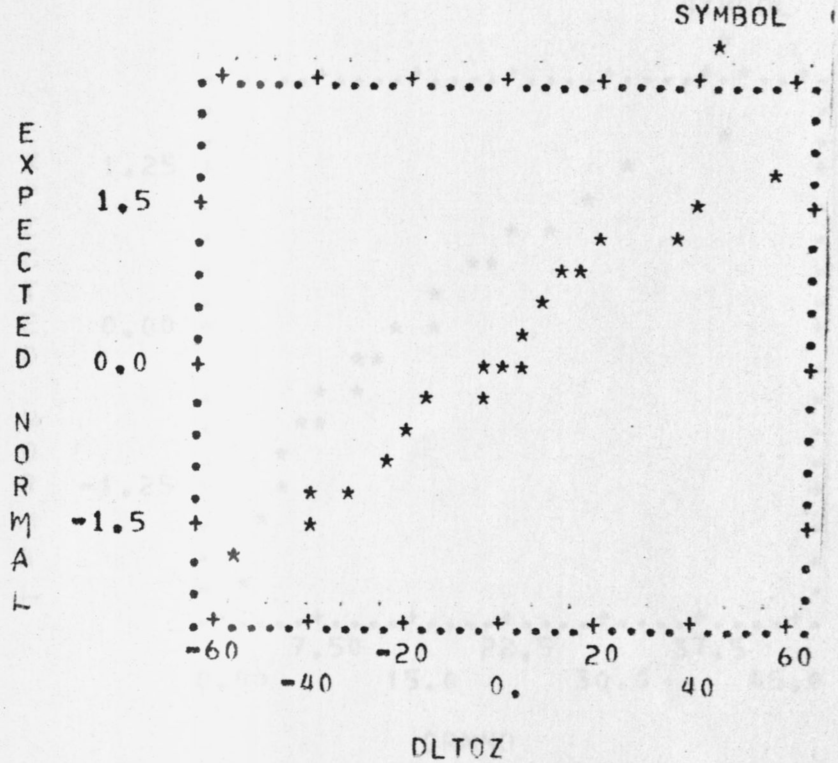
NORMAL PLOT OF VARIABLE 9 GRNOZ




```
INTERVAL
NAME            5    10
              +----+----+-
* 32.000 +X
* 40.000 +XXX
* 48.000 +X
* 56.000 +XXX
* 64.000 +XXXXX
* 72.000 +XX
* 80.000 +
* 88.000 +XXXX
* 96.000 +XX
* 104.00 +XX
* 112.00 +X
* 120.00 +XX
* 128.00 +
              +----+----+-
                   5    10
```

N = 26
X = 70.5
s = 25.8
CV = .37
skew = .21
kurt = −1.03
min = 24
max = 117

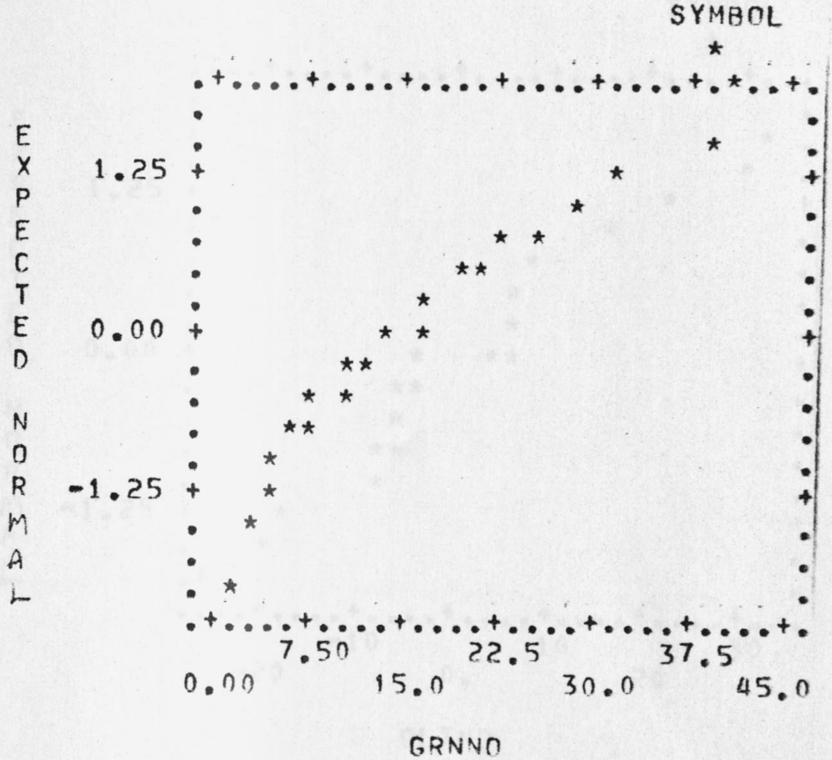
NORMAL PLOT OF VARIABLE 15 DLTOZ




```
INTERVAL
NAME             5    10
              +----+----+-
*-50.000  +X
*-40.000  +XX
*-30.000  +X
*-20.000  +XXXX
*-10.000  +X
* 0.0000  +XXXX
* 10.000  +XXXXXX
* 20.000  +XXX
* 30.000  +
* 40.000  +XX
* 50.000  +
* 60.000  +X
* 70.000  +
              +----+----+-
                   5    10
```

N = 25
X = -3.5
s = 25.9
CV =
skew = .10
kurt = -.32
min = -57
max = 55

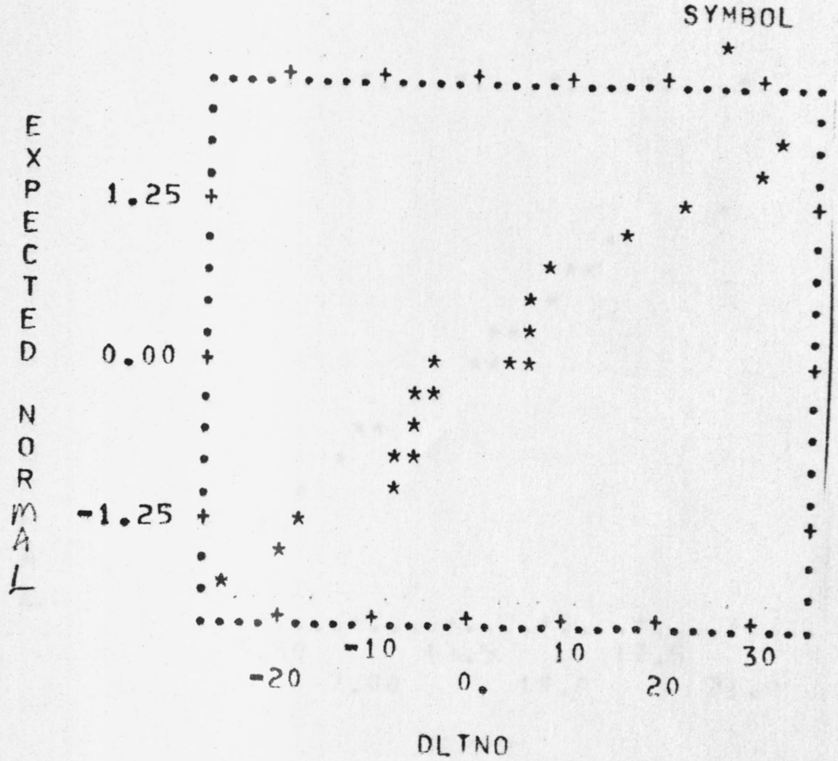
NORMAL PLOT OF VARIABLE 10 GRNNO




```
INTERVAL
NAME              5    10
             +----+----+-
* 7.0000     +XXXXXXX
* 10.500     +X
* 14.000     +XXX
* 17.500     +XXX
* 21.000     +XX
* 24.500     +X
* 28.000     +X
* 31.500     +X
* 35.000     +X
* 38.500     +
* 42.000     +XX
* 45.500     +
* 49.000     +
             +----+----+-
                  5    10
```

N = 18
X = 16.3
s = 11.5
CV = .70
skew = .80
kurt = -.53
min = 3
max = 41

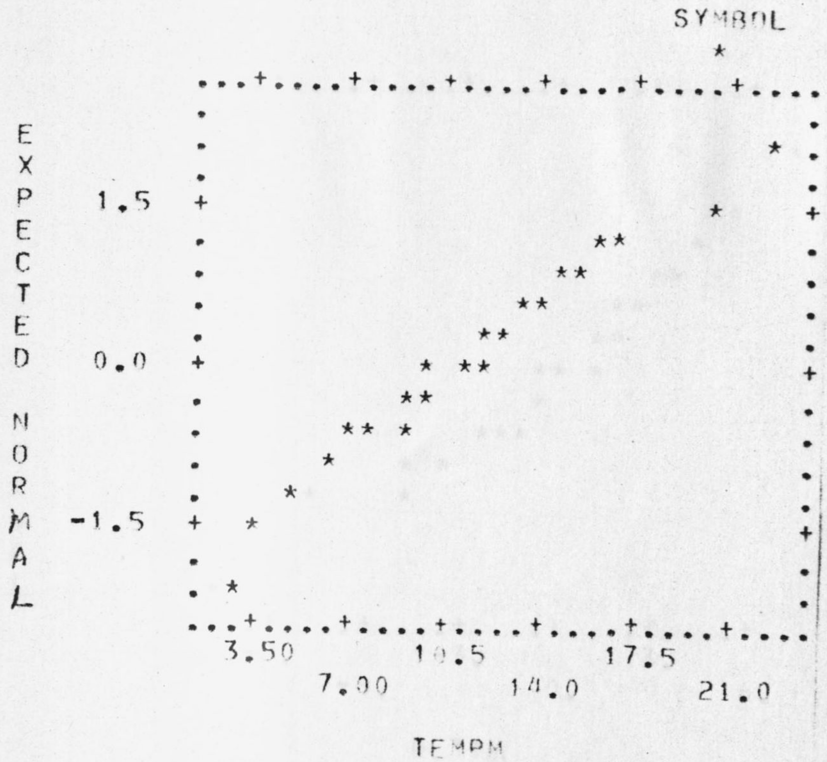
NORMAL PLOT OF VARIABLE 16 DLTNO




```
INTERVAL
NAME            5    10
             +----+----+-
*-20.000  +X
*-15.000  +XX
*-10.000  +
*-5.0000  +XXXXXXX
* 0.0000  +
* 5.0000  +XXX
* 10.000  +XXXX
* 15.000  +
* 20.000  +X
* 25.000  +X
* 30.000  +X
* 35.000  +X
* 40.000  +
          +----+----+-
               5    10
```

N = 18
X = 1.9
s = 15.2
CV =
skew = .38
kurt = −.68
min = −25
max = 32

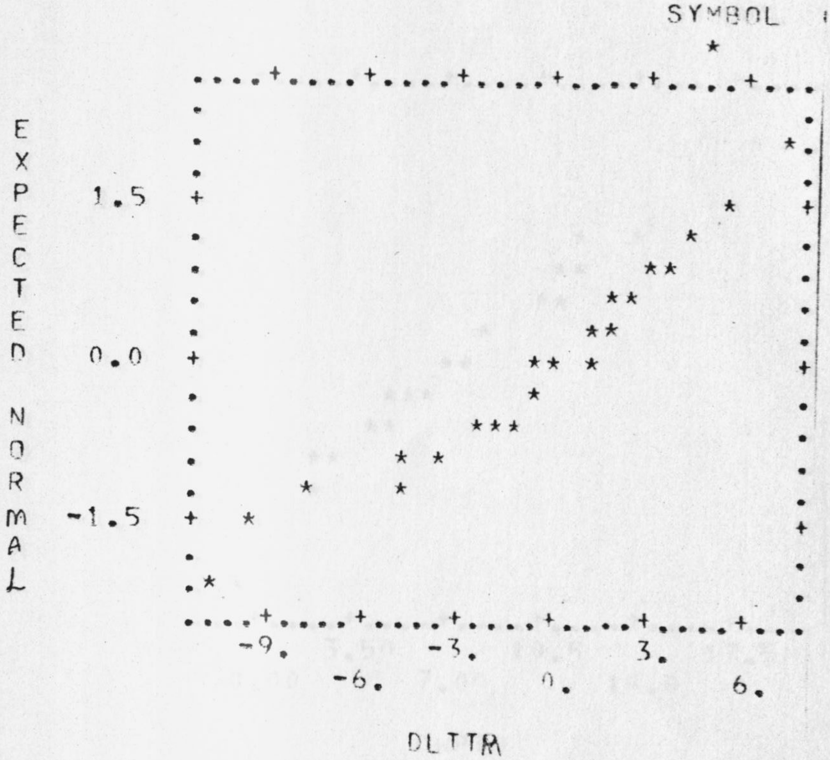
NORMAL PLOT OF VARIABLE 5 TEMPM




```
INTERVAL
NAME              5    10
            +----+----+-
* 3.6000    +XX
* 5.4000    +XX
* 7.2000    +XXX
* 9.0000    +XXX
* 10.800    +XXX
* 12.600    +XXXXX
* 14.400    +XXXX
* 16.200    +XX
* 18.000    +XX
* 19.800    +
* 21.600    +X
* 23.400    +X
* 25.200    +
            +----+----+-
                 5    10
```

N = 26
X = 11.4
s = 5.0
CV = .41
skew = .25
kurt = −.21
min = 2.9
max = 22.2

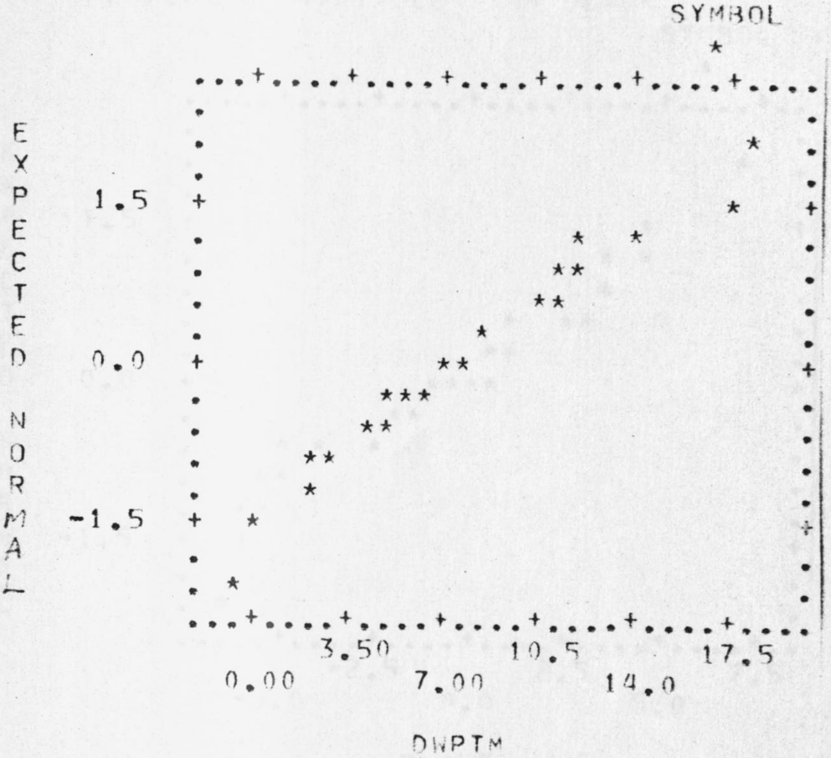
NORMAL PLOT OF VARIABLE 7 DLTTM




```
INTERVAL
NAME              5    10
              +----+----+-
*-9.0000  +XX
*-7.5000  +X
*-6.0000  +
*-4.5000  +XX
*-3.0000  +X
*-1.5000  +XX
* 0.0000  +XXXXXX
* 1.5000  +XXX
* 3.0000  +XXXXXX
* 4.5000  +XXX
* 6.0000  +X
* 7.5000  +X
* 9.0000  +
              +----+----+-
                   5    10
```

N = 25
X = -0.7
s = 4.4
CV =
skew = .53
kurt = -.30
min = -10.5
max = 7.1

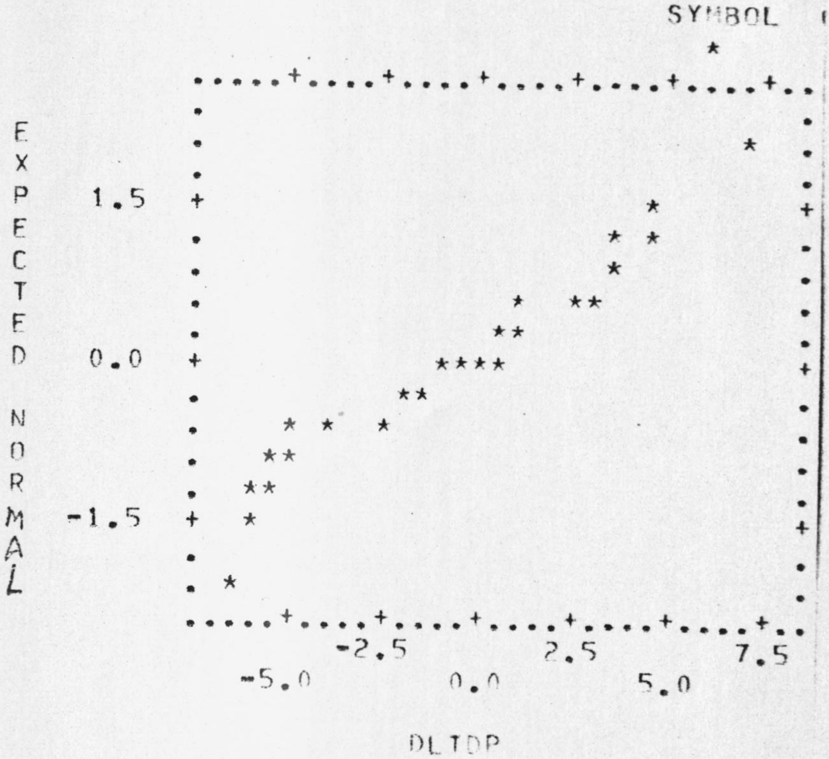
NORMAL PLOT OF VARIABLE 6 DWPTM




```
INTERVAL
NAME            5    10
             +----+----+-
* 1.8000     +XX
* 3.6000     +XXXX
* 5.4000     +XXXX
* 7.2000     +XXXX
* 9.0000     +XXXXX
* 10.800     +XX
* 12.600     +XXXX
* 14.400     +X
* 16.200     +
* 18.000     +X
* 19.800     +X
* 21.600     +
* 23.400     +
             +----+----+-
                  5    10
```

N = 26
X = 7.9
s = 4.8
CV = .61
skew = .29
kurt = −.47
min = −.6
max = 18.1

NORMAL PLOT OF VARIABLE 8 DLTDP




```
INTERVAL
NAME              5    10
          +----+----+-
*-4.8000  +XXXXXXX
*-3.6000  +X
*-2.4000  +X
*-1.2000  +XXX
* 0.0000  +XXX
* 1.2000  +XXXXX
* 2.4000  +
* 3.6000  +XXXXX
* 4.8000  +XX
* 6.0000  +
* 7.2000  +X
* 8.4000  +
* 9.6000  +
          +----+----+-
               5    10
```

N = 18
X = -.28
s = 3.9
CV =
skew = -.47
kurt = -.42
min = -6.2
max = 7.1