DOCUMENT RESUME

ED 082 469 EM 011 428

AUTHOR Munick, Herman; Allison, John
TITLE On Uses and Misuses of Computer Programs in Statistics.
PUB DATE Jun 73
NOTE 9p.; Paper presented at the Conference on Computers in the Undergraduate Curricula (4th, Claremont, California, June 18-20, 1973)

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS *Computer Assisted Instruction; *Computer Programs; Correlation; *Statistical Analysis; *Statistics
IDENTIFIERS Autocorrelation; Distributions; Linear Regressions; Normal Curve; Normal Distribution; T Values

ABSTRACT

Distributions and linear regressions are discussed. The section dealing with the former topic emphasizes the usefulness of computer programs in statistics, demonstrating their ability to handle tedious and time-consuming tasks. The normal curve is stressed since the assumption of a normal distribution is common. An example of 200 data points is presented which illustrates a computer program's ability to do descriptive statistics for data not grouped, data which is distributed into any number of classes of equal width and testing if the 200 points are normally distributed. The section of linear regression emphasizes how computer programs can lead to erroneous results. In the area linear regression there are in current use "canned programs" which display the t-values corresponding to the coefficients of the least square estimate, but do not take autocorrelation into account. In the presence of autocorrelation it is incorrect to use these t-values. An example is presented in which autocorrelation is present but can be removed by a suitable transformation of variables. (Author)

ON USES AND MISUSES OF COMPUTER PROGRAMS IN STATISTICS

Dr. Herman Munick*
College of Business
St. John's University
Jamaica, New York 11432
(212) 969-8000

John Allison
940 East 27th Street
Brooklyn, New York 11210
(212) 252-7096

This paper is divided into two sections, DISTRIBUTIONS, and LINEAR REGRESSION. The first section DISTRIBUTIONS emphasizes the usefulness of computer programs in statistics. Some very tedious and time-consuming tasks are made considerably easier and clarified by use of these computer programs. In particular, the normal curve is emphasized since in many areas of statistics the assumption "a population is normally distributed" appears time and time again. An example of 200 data points is presented illustrating the computer program's ability to do descriptive statistics for data not grouped, data which is grouped into any number of classes of equal width, and testing if the 200 points are normally distributed. A graph of the data also appears in the output.

The second section LINEAR REGRESSION emphasizes how computer programs can lead to erroneous results. In the area of linear regression there are in current use "canned programs" which display the t-values corresponding to the coefficients of the least square estimate, but do not take autocorrelation into account. In the presence of autocorrelation it is incorrect to use these t-values. An example of this is presented where autocorrelation is present but can be removed by a suitable transformation of variables.

Distributions

The following are the salient features of a program written to obtain the following:

1. Descriptive statistics for data which is not-grouped.
2. Indicators that a set of points are normally distributed (necessary but not sufficient conditions).
3. Grouping original data into any number of classes of equal width, with descriptive statistics.
4. A graph of the grouped data (frequency plotted against mid-point of CLASS).
5. CHI-SQUARE test for goodness of fit of normal curve.

Example

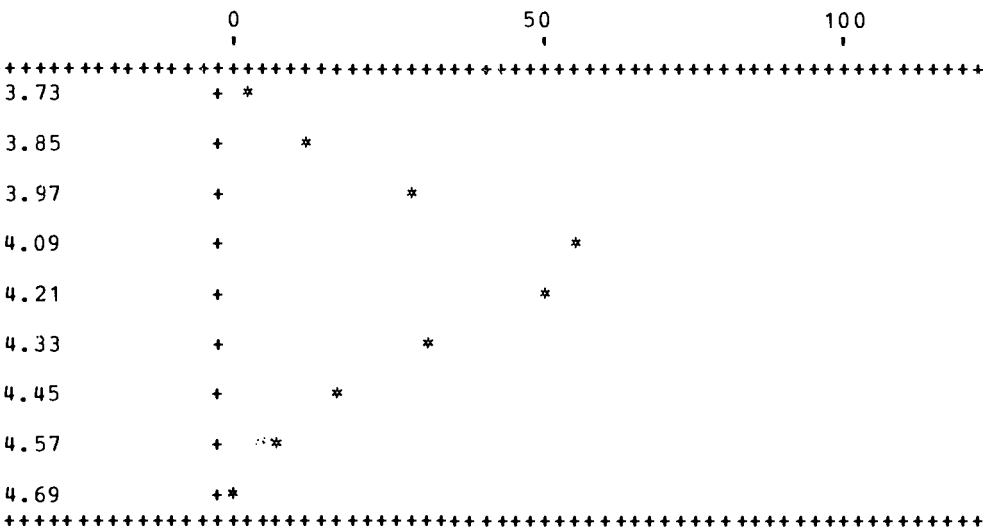
The following is a display of an input and output of the program for data points as indicated:

```
9900 DATA 4.7,4.62,4.48,4.45,4.2,4.14,4.1,4.08,4,4.39
9901 DATA 4.44,4.35,4.29,4.28,4.13,4.33,4.37,4.43,4.59,4.1
9902 DATA 4.49,4.34,4.33,4.24,4.2,4.17,4.28,3.76,4.16,4.5
9903 DATA 4,4.07,4.22,4.53,4.55,4.44,4.02,4.2,4.13,4.16
9904 DATA 4.2,4.13,4.12,4.19,4.18,3.98,3.96,4.17,4.2,3.99
9905 DATA 4.16,4.44,4.33,4.59,4.54,4.16,4.21,4.42,4.33,4.34
9906 DATA 4.13,4.31,4.22,4.33,4.12,4.46,4.27,4.21,4.2,4.05
9907 DATA 4.26,4.15,4.08,4.12,4.1,4.15,4.07,4.25,4.3,4.29
9908 DATA 4.45,4.37,4.2,4.31,4.09,4.23,4.2,4.19,4,4.39
9909 DATA 4.38,4.48,4.42,4.38,4.11,4.13,3.68,4.23,4.23
9910 DATA 4.17,4.25,4.18,4.19,3.99,4.07,3.76,4.12,4,4.09
9911 DATA 3.34,4.13,3.37,4.21,4.19,4.28,4.02,4.02,4.24,4.07
9912 DATA 4.35,4.27,4.05,4.14,4.16,4.13,4.19,4.05,3.91,3.89
9913 DATA 4,4.03,4.07,3.98,4.2,4.1,4.13,4.07,4.28,4.06
9914 DATA 4.45,4.24,4.08,4.09,3.95,3.95,3.87,4.13,4.09,3.97,3.95
9915 DATA 3.93,4.25,4.23,3.99,3.95,4,3.95,4.08,4.13,4.13
9916 DATA 4.36,4.36,4.06,4.2,4.17,3.88,3.99,3.98,4.1,3.88
9917 DATA 4.02,4.18,4.35,4.08,4.06,3.92,3.9,4.16,4.14,4.06
9918 DATA 4.25,4.41,4.07,4.08,4.04,4.1,4.21,3.87,3.9,3.93
9919 DATA 4.15,4.08,4.29,4.07,4.01,4.25,4.03,4.27,4.07,4.17
9998 DATA -1
9999 END
```

```
CHOOSE WHAT INFORMATION YOU NEED

     A) NOT-GROUPED DATA
     B) GROUPED DATA
     C) BOTH
     D) NONE OF THE ABOVE

ANSWER A,B,C, OR D
?C
HOW MANY CLASSES ARE THERE?
?9
WHAT IS THE MINIMUM NUMBER OF THE FIRST CLASS?
?3.67
WHAT IS THE MAXIMUM NUMBER OF THE LAST CLASS?
?4.75

                    STATISTICS FOR 'NOT-GROUPED' DATA

          THERE WERE 200 OBSERVATIONS MADE
          THE AVERAGE IS 4.1714
          THE VARIANCE IS 3.09021E-02
          THE STANDARD DEVIATION IS .17579
          THE MAXIMUM VALUE IS 4.7
          THE MINIMUM VALUE IS 3.68
          THE RANGE OF THE OBSERVATIONS IS 1.02
          THE MEASURE OF SYMMETRY IS .260124
          THE MEASURE OF PEAKEDNESS IS 3.10359


          THE FIRST DEVIATION BAND BETWEEN 3.99561     AND 4.34719
               CONTAINS 68% OF THE OBSERVATIONS

          THE SECOND DEVIATION BAND BETWEEN 3.81982    AND 4.52298
               CONTAINS 95% OF THE OBSERVATIONS

          THE THIRD DEVIATION BAND BETWEEN 3.64403     AND 4.69877
               CONTAINS 99.5% OF THE OBSERVATIONS
```

STATISTICS FOR 'GROUPED' DATA

CLASS BOUNDARIES FROM	TO	MIDPOINT	FREQUENCY
3.67	3.79	3.73	3
3.79	3.91	3.85	9
3.91	4.03	3.97	28
4.03	4.15	4.09	54
4.15	4.27	4.21	51
4.27	4.39	4.33	31
4.39	4.51	4.45	17
4.51	4.63	4.57	6
4.63	4.75	4.69	1



THE NUMBER OF OBSERVATIONS MADE WERE 200
THE AVERAGE IS 4.1728
THE MODE IS 4.09
THE VARIANCE IS 3.20242E-02
THE STANDARD DEVIATION IS .178953
THE MEDIAN IS 4.16412

O(I)	E(I)	(O(I)-E(I))²/E(I)
3	2.568	.072673
9	10.696	.268924
28	28.416	6.09010E-03
54	48.192	.699968
51	52.202	2.76773E-02
31	36.108	.722601
17	15.958	6.80388E-02
6	4.498	.501557
1	.8088	4.51996E-02
TOTAL		2.41273 (Computed Chi-square value)

Analysis of computer output:

1. The usual measures such as average, variance, and standard deviation are obtained. In addition the maximum value, minimum value and range are determined.

2. Preliminary indications are that the set of 200 numbers is normally distributed since the measure of symmetry is approximately zero, peakedness is approximately 3, and one, two, and three standard deviation bands contain respectively 68%, 95%, and 99.5% of the data.

3. Nine classes were asked for and the grouped data for these nine classes is presented with the midpoint of each class. The usual measures such as average, variance, standard deviation, median and mode are given for the data grouped into these nine classes. It is noted that the average and standard deviation for the grouped data (4.1728 and .178953) are approximately the same for the data when not-grouped (4.1714 and .17579).

4. The graph indicates the data is approximately symmetric about the average value of 4.1714.

5. The CHI-SQUARE value of 2.41273 indicates that the normal curve (average of 4.11728 and standard deviation of .178953) fit is good at both the 5 percent and 1 percent level of significance (O(I) is observed frequency and E(I) is expected frequency).

Conclusions:

The program is particularly valuable in that it will offer descriptive statistics for any number of classes asked for. In some realistic situations it is important to compare results using different numbers of classes and this can now be done with considerable ease. The graph of the data is of particular importance in that one can quickly get an initial impression of how the data behaves. The program emphasizes the normal distribution. The reason for this is that in such topics as linear regression (second section of this paper) an assumption often stated is that a set of points is normally distributed.

Linear Regression

An assumption in the linear regression model is that the error terms are statistically independent. If violated, it is referred to as a problem of autocorrelation, invalidating use of the usual t-tests. There are available computer "canned programs" currently which display the t-values corresponding to the coefficients of the least square estimate, but do not take autocorrelation into account. Therefore, in the presence of autocorrelation it is incorrect to use these t-values [2, p. 80].

Under suitable conditions, depending on the given data, autocorrelation can be eliminated by transformation to a new set of variables. In the transformed system it is valid to use the t-values. This subject is treated clearly by Frank [3]. A typical case where autocorrelation is not considered is Parker and Segura's article [4], an interesting and useful paper. In their article, sales are forecast in the home furnishing industry using new marriages during the year, housing starts, annual disposable income, and time trend as explanatory variables, using 24 years of data. This is an example of time-series data where one must be especially careful for the presence of autocorrelation. In the transformed system it is demonstrated that autocorrelation is eliminated and the least square estimate provides a reasonably good estimate of the original set of data with the t-values significant at approximately a 5% level of significance.

GIVEN 24 YEARS OF DATA

Year	Housing starts (H) (thousands)	Disposable personal income (I) ($ billions)	New marriages (M) (thousands)	Company sales ($) ($ millions)	Time (T)
1947	744	158.9	2,291	92.920	1
1948	942	169.5	2,991	122.440	2
1949	1,033	188.3	1,811	125.570	3
1950	1,138	187.2	1,580	110.460	4
1951	1,549	205.8	1,667	139.400	5
1952	1,211	224.9	1,595	154.020	6
1953	1,251	235.0	1,539	157.590	7
1954	1,225	247.9	1,546	152.230	8
1955	1,354	254.4	1,490	139.130	9
1956	1,475	274.4	1,531	156.330	10
1957	1,240	292.9	1,585	140.470	11
1958	1,157	308.5	1,518	128.240	12
1959	1,341	318.8	1,451	117.450	13
1960	1,531	337.7	1,494	132.640	14
1961	1,274	350.0	1,527	126.160	15
1962	1,274	364.4	1,547	116.990	16
1963	1,469	385.3	1,580	123.900	17
1964	1,615	404.6	1,654	141.320	18
1965	1,538	436.6	1,719	156.710	19
1966	1,488	469.1	1,789	171.930	20
1967	1,173	505.3	1,844	184.790	21
1968	1,299	546.3	1,913	202.700	22
1969	1,524	590.0	2,059	237.340	23
1970	1,479	629.6	2,132	254.930	24

Define the following for year i:

H_i = Housing Starts
I_i = Annual Disposable Income
M_i = New Marriages
S_i = Gross Sales
T_i = Time Trend
$i = 1,2,\ldots,24$

Presence of autocorrelation in error terms

Using a linear regression computer program the following least square estimate is obtained:

$$S = 50.605 + .036H + 1.221I - .068M - 19.483T \qquad (1)$$

The following table is then established:

YEAR	GROSS SALES	ESTIMATED GROSS SALES	ERROR TERMS
1	92.940	95.673	-2.752
2	122.440	116.729	5.711
3	125.570	135.752	-10.181
4	110.460	134.467	-24.006
5	139.400	146.578	-7.178
6	154.020	143.129	10.891
7	157.590	141.237	16.354
8	152.230	136.085	16.145
9	139.130	133.008	6.122
10	156.330	139.506	16.824
11	140.470	130.447	10.024
12	128.240	131.581	-3.341
13	117.450	135.877	-18.427
14	132.640	143.384	-10.744
15	126.160	127.395	-1.236
16	116.990	126.037	-9.047
17	123.900	134.937	-11.037
18	141.320	139.232	2.088
19	156.710	151.599	5.111
20	171.930	165.210	6.720
21	184.790	174.802	9.988
22	202.700	205.204	-2.504
23	237.340	238.319	-.979
24	254.930	259.474	-4.544

The t-values corresponding to the coefficients of the least square estimate are not displayed here since it will be shown that there is the presence of first order autocorrelation invalidating the use of these t-values. The main point of this section is that a computer program is incorrect if it displays these t-values if autocorrelation is present.

Denoting by $\hat{\ell}_i$ the error term for year i , a test for first order autocorrelation is based on the Durbin-Watson statistic [3, p. 276].

$$d = \frac{\sum_{i=2}^{24} \left(\hat{\ell}_i - \hat{\ell}_{i-1}\right)^2}{\sum_{i=2}^{24} \hat{\ell}_i^2} = .8852 \qquad (2)$$

Since the number of observations is 24 and 4 independent variables are under consideration it follows from Table E of Frank's book [3, p. 276] that the d_1 and d_2 for a two-tailed critical region of 5 per cent are given by

$$d_1 = .91 \qquad d_2 = 1.66 \qquad (3)$$

and that therefore

$$d < d_1 < d_2 \qquad (4)$$

Concluding that there is significant first-order autocorrelation [3, p. 279].

Transformation to a new set of variables--eliminating autocorrelation

The first step is to determine ρ, the coefficient of autocorrelation. The estimated value of ρ is given by Frank [3, pp. 280-281].

$$\hat{\rho} = \frac{\sum_{i=2}^{24} \hat{\ell}_i \, \hat{\ell}_{i-1}}{\sum_{i=2}^{24} \hat{\ell}_i^2} \qquad (5)$$

An approximation to $\hat{\rho}$ is given by

$$\hat{\rho} = 1/2\ (2-d) = .5574 \qquad (6)$$

where d is the Durbin-Watson statistic.

Denoting by H_i, I_i, M_i, S_i, T_i a new set of variables the transformation is given by [Frank [3, p. 280]

$$\begin{aligned} H_i &= H_i - .5574\ H_{i-1} \\ I_i &= I_i - .5574\ I_{i-1} \\ M_i &= M_i - .5574\ M_{i-1} \\ S_i &= S_i - .5574\ S_{i-1} \\ T_i &= T_i - .5574\ T_{i-1} \end{aligned} \qquad (7)$$

where i = 2,3, . . . ,24.

Performing these transformations leads to the following set of columns:

i	H_i'	I_i'	M_i'	S_i'	T_i'
1	--	--	--	--	--
2	527.303	80.931	714.022	70.647	1.443
3	507.940	93.823	701.239	57.323	1.885
4	562.217	82.244	570.568	40.469	2.328
5	914.691	101.457	786.325	77.831	2.770
6	347.604	110.189	665.833	76.320	3.213
7	576.002	109.643	649.964	71.741	3.656
8	527.706	116.914	688.178	64.391	4.098
9	671.198	116.223	628.277	54.279	4.541
10	720.295	132.600	700.490	78.781	4.984
11	417.851	139.952	731.637	53.333	5.426
12	465.838	145.241	634.538	49.944	5.869
13	696.101	146.546	604.883	45.970	6.311
14	783.541	160.172	685.229	67.175	6.754
15	420.637	161.770	694.261	52.228	7.197
16	616.886	169.314	695.867	46.670	7.639
17	729.345	182.187	717.719	58.691	8.082
18	796.196	189.838	773.325	72.260	8.524
19	637.817	211.080	797.079	77.940	8.967
20	630.736	225.744	830.848	84.582	9.410
21	343.605	243.829	846.831	88.958	9.852
22	645.183	264.651	885.175	99.700	10.295
23	799.952	286.398	992.715	124.357	10.737
24	629.539	300.239	984.336	122.639	11.180

For these transformed variables the least square estimate is given by

$$S' = -43.879 + .023H' + .713I' + .088M' - 12.709T' \qquad (8)$$

In the transformed system the error terms are given as follows:

i	S_i	S_i (ESTIMATED)	ERROR TERMS
1	--	--	--
2	70.647	70.590	.057
3	57.323	72.583	-15.260
4	40.469	48.479	-8.010
5	77.831	83.754	-5.923
6	76.320	60.525	15.795
7	71.741	58.447	13.294
8	64.391	60.240	4.152
9	54.279	52.205	2.074
10	78.781	65.755	13.026
11	53.333	61.051	-7.718
12	49.944	51.782	-1.838
13	45.970	49.856	-3.885
14	67.175	63.053	4.121
15	52.228	50.890	1.337
16	46.670	55.368	-8.698
17	58.691	63.470	-4.779
18	72.260	69.751	2.508
19	77.940	77.667	.273
20	84.582	85.304	-.722
21	88.958	87.279	1.679
22	99.700	106.915	-7.215
23	142.357	129.868	12.489
24	122.639	129.398	-6.759

For these error terms the Durbin-Watson statistic is given by

$$d = 1.8015 \qquad (9)$$

Since 23 observations and 4 independent variables are under consideration one establishes from Table E of Frank's book [3, p. 276] that

$$d_1 = .89 \qquad d_2 = 1.67 \qquad (10)$$

And therefore in the transformed system

$$d > d_2 > d_1 \qquad (11)$$

in which case one accepts the hypothesis of no first order autocorrelation [3, p. 279]. Therefore in the transformed system it is valid to use the t-values.

Significance of t-values

The transformed equation with the corresponding t-values below them are

$$\begin{array}{l} S = -43.879 + .023H + 713I + 088M - 12.709T \\ \quad\quad -1.952 \quad 1.694 \quad 2.959 \quad 1.658 \quad 3.237 \end{array} \qquad (12)$$

Coefficient of Determination = .893

Standard Error of Estimate = 8.845

In order to determine the significance of the t-values one uses a Student's-t disbribution with (N-K) degrees of freedom. Since in the transformed system

N = number of observations = 23
K = total number of variables under consideration = 5

It follows that the number of degrees of freedom is 18. The critical value for a 5% one-tailed test of significance using the Student's-t distribution with 18 degrees of freedom is 1.729. The t-values associated with the constant term, dispcsable personal income and time trend are significant at a 5% level. The t-values asscciated with new housing starts and new marriages are significant at approximately a 5% level.

NOTES AND REFERENCES

1. Ya-Lun Chou, Statistical Analysis. Holt, Rinehart, and Winston, Inc., 1969.

2. W. L. Hays and L. W. Winkler, Statistics, Vol. II. Holt, Rinehart and Winston, Inc., 1970.

3. C. R. F. Frank, Jr., Statistics and Econometrics. Holt, Rinehart and Winston, Inc., 1971.

4. G. C. Parker and E. L. Segura, How to get a better forecast. Harvard Business Review, 1971, March-April, 99-109.

*Please send correspondence to Dr. Munick.