11-1978

Use of Correlation and Regression Analysis to Determine Bus Operations

Murray Frost
University of Nebraska at Omaha

THE USE OF CORRELATION AND REGRESSION ANALYSES TO DETERMINE BUS OPERATOR HOURLY WAGES

by Dr. Murray Frost

Center for Applied Urban Research
The University of Nebraska at Omaha
Omaha, Nebraska 68182

November 1978

Introduction

The purpose of this discussion is to assess the use of correlation and regression analyses for determining bus operator hourly wages. If these hourly wages in a set of cities are highly correlated with another variable, it becomes possible to predict these wages by measuring this other variable.

The data used here are those presented to the Nebraska Court of Industrial Relations by the Transport Workers Union (TWU) related to its contract negotiations with Metro Area Transit (MAT) which operates the transit system in the Omaha metropolitan area.

The data presented to the Center for Applied Urban Research for analysis represented bus operator hourly wages in 11 cities including Omaha. The first factor that must be noted is the selection of the cities. The North Central region according to the Bureau of the Census includes Ohio, Indiana, Illinois, Michigan, and Wisconsin in the East North Central part of the region, and Minnesota, Iowa, Missouri, North Dakota, South Dakota, Nebraska, and Kansas in the West North Central part of the region. These states have 41 cities with over 100,000 population. The population of the 11 selected cities ranges from 149,518 to 507,242. No reasons are offered for excluding the seven cities larger than 507,242 nor for excluding the 16 cities below 149,518 in the region. More importantly, no reason is given for excluding seven cities in the region within this population range: Flint (193,000), Fort Wayne (178,000), Gary (175,000), Grand Rapids (198,000), Kansas City, Kansas (168,000), Madison (173,000), or Warren (179,000). Bus operator hourly wage data are apparently available for at least five of these cities (see p. 52 of the data submitted).

The second factor which is relevant is the selection of the variables which are measured in relation to bus operator hourly wages. No explanation is offered for the 11 variables selected: 1970 and 1973 population, population density, per capita and median family income, workers using transit to work, mean January and July temperatures, annual precipitation, percent employment in manufacturing, and percent of unionization of plant

workers.

Correlation Analysis

The third consideration relates to the statistics used to analyze the data--specifically, simple correlation analysis which is a measure of the association of two variables. An essential first step in using correlation (or regression), which is a parametric statistic, is to determine whether the data meet the assumptions necessary to use these statistics. These assumptions include: 1) linear relationship between the two variables--i.e., for every unit change in the Y variable there is a constant unit change in the X variable (e.g. drivers' hourly wages may increase 20¢ for each percentage point increase in workers using transit), and 2) homoscedascity of the variables--i.e. the X variable must not have a low variance for some values of Y and a high variance for other values of Y. (In other words, the spread of points around the regression line should be about the same all along the line.)

Whether or not the data fit these assumptions can be determined at least approximately by examining a scatter diagram of the data points. Several of these scatter diagrams indicate that the data for some pairs of variables do not meet the assumptions of linearity and homoscedascity (see Appendix 2).

In the definitional formula for correlation the coefficient of correlation (r) equals the sum of the product of the deviations for the X and Y variables from their respective means divided by the product of the standard deviations for each variable. The coefficient of correlation will vary between +1 and -1. (If there is a perfect direct linear relationship, it will be +1; if there is a perfect inverse linear relationship, it will be -1; and if there is no linear relationship, it will be 0).

The formula used for computations is:

$$r = \frac{\Sigma XY - \frac{(\Sigma X)\ (\Sigma Y)}{N}}{\sqrt{\left[\Sigma X^2 - \frac{(\Sigma X)^2}{N}\right]\left[\Sigma Y^2 - \frac{(\Sigma Y)^2}{N}\right]}}$$

Once the correlation coefficient is computed, a test to determine its statistical significance (or reliability) should be made. This test does not address the value judgment of the utility of the data.* It is merely a measure of the probability of the correlation coefficient occurring by chance. Social scientists generally operate at the .05 confidence level--i.e. they are willing to be wrong 5% of the time (or willing to be right only 95% of the time).

The statistical significance of a correlation coefficient depends in part on the number of observations used to calculate it. If a sample of 41 cities is used, a correlation coefficient is less likely to have been the result of chance than if the same correlation coefficient had been calculated on the basis of 11 cities. In other words, the more observations involved, the smaller the correlation coefficient can be and still be significant (not produced by chance).

A formula to determine the critical value of a significant correlation coefficient is:

$$r_{crit} = \sqrt{\frac{t^2}{t^2 + N - 2}}$$

If the 5% (or 95%) confidence level is acceptable, then for 11 sets of observations the correlation coefficient must be .60 or higher to be considered significant; for only seven sets of observations the correlation coefficient must be .75 to be considered "significant."

If a correlation coefficient is given, its statistical significance--as measured by t--can be determined from the following formula:

$$t = \left[\frac{r}{\sqrt{1-r^2}}\right]\left[N-2\right] \quad \text{with N-2 degrees of freedom}$$

The correlation coefficient can also be used to estimate the amount of variation in Y that is determined (or "explained") by the variation in X. This is done by squaring the correlation coefficient. Thus a correlation coefficient of .75 between bus drivers' hourly wages (Y) and the proportion of workers using transit (X), "explains" only 56% of the

*See Abraham N. Franzblou, A Primer of Statistics for Non-Statisticians (New York: Harcourt, Brace & World, Inc., 1958), pp. 76-9.

variation of bus drivers' hourly wages. If this correlation coefficient is .60, it "explains" only 36% of the variation.

An examination of the correlation coefficients for the association of bus operators' hourly wages and each of the other variables obtained from the data for the 11 cities (including Omaha) submitted by the TWU to the Court of Industrial Relations indicates only two correlations significant at the .05 level or better. Bus operators' hourly wages correlates at .747 with the percentage of workers using transit, and at ,617 with the cities' 1970 population.*

As one social science text on quantitative techniques notes, "A correlation coefficient measures the degree of association between two sets of paired variates. A significance test may show whether that degree of association is likely to be more than a matter of chance. What neither the correlation coefficient nor a significance tells us is the way in which the two sets of variates are related: they cannot be used to predict one set of variates from a knowledge of the other."** Regression analysis is the statistical tool used to pursue that goal.

Regression Analysis

Simple regression analysis is used to predict one variable on the basis of another. In essence regression tells how much better an estimate of a value of the Y variable is if it is based on a measurement of the X variable than if it were based on the mean of Y variable observations (e.g. how much better one can predict Omaha's bus operators' hourly wages by measuring the proportion of workers using transit than by using the average of bus drivers' hourly wages in a number of cities). Multiple regression analysis is used to predict a variable on the basis of two or more other variables.

The assumptions are the same for regression as for correlation, and the caveat that regression does not demonstrate a causal relationship must still be made (that one can predict a value of Y by measuring X does

---

*The latter is based on the correct 1970 population for Omaha (347,380) not the population provided in the report (p. 8 reports the population as 247,380). It should be noted that a later population estimate (1973) was not correlated significantly (at the .05 level) with bus drivers' hourly wages in July, 1978.

**Robert Hammond and Patrick McCullagh, Quantitative Techniques in Geography: An Introduction (Oxford: Clarendon Press, 1974), p. 218.

not mean that X causes Y--they may both be caused by another variable, or the relationship may be spurious, i.e. there is no logical connection between the two). The assumptions of homoscedascity and linearity can be tested approximately by examining a scatter diagram for each pair of X and Y variables. Regression, therefore, is a way of drawing a line through a scatter diagram that best summarizes the scatter.

The formula for such a line takes the form of $Y = bx + c$, where b is the regression coefficient and c is the point where that line intercepts the Y (or vertical axis) of the scatter diagram.

The computational formula for $$b = \frac{\Sigma XY - \frac{(\Sigma X)(\Sigma Y)}{N}}{\Sigma X^2 - \frac{(\Sigma X)^2}{N}}$$

and for $c = \bar{y} - b\bar{x}$.

Once b and c are calculated a value of Y can be predicted on the basis of the value of X. In other words, if one wanted to estimate a city's bus drivers' hourly wages (Y) on the basis of the proportion of workers using transit in that city, then b and c could be calculated from observations of these variables in a number of cities.

Using the 11 city data provided to CAUR, the regression coefficient (b) for predicting bus drivers' hourly wages based on the proportion of workers using transit was .137 and the intercept (c) was 4.904. Therefore the calculation for predicting Y in any city is: $Y = .137\ (X) + 4.904$.

Since the percent using transit (X) in Omaha is 9.3, the predicted wage in Omaha would be: .137 (9.3) + 4.904 or 1.274 + 4.904 or $6.17. This is actually less than the $6.24 Omaha pays its drivers.*

But this estimate would be accurate only if the relationship between the two variables were proportional--i.e. the data points fell on a straight line. But since this is not the case, as an examination of the

*This estimate is based on computations using a regression coefficient (beta) accurate to 3 decimal places. The estimate may vary if the regression coefficient is carried out to more digits. For example the estimate of Omaha drivers' hourly wages based on a measurement of median income is $10.30 if the regression coefficient is accurate to 3 places, $6.22 if it is accurate to 4 places, and is $6.29 if it is carried out to 6 places. The user, therefore, should be aware of this rounding error when using these statistics.

scatter diagram indicates, an estimate of confidence in the prediction should be calculated. The standard error of the estimate (S) is equal to the square root of the squared differences between observed and expected Y values, divided by N - 2. The computational formula is:

$$S = \sqrt{\frac{\Sigma Y^2 - \frac{(\Sigma XY)^2}{\Sigma X^2}}{n-2}}$$

As can be seen, the use of a small number of observations (11) increases the size of the error. Many authors claim that the standard error of the estimate should always be given along with regression predictions.*

For this particular example the standard error of the estimate is .709. Therefore the predicted value will be within $\pm$ .709 of the real value 68% of the time and within $\pm$ 1.42 95% of the time. This wide confidence limit undermines the validity or utility of the predictions. It is important to note that the error of .709 or 71¢ is larger than the difference which is in dispute, which is only about 19¢ (the difference between 6% and 9% of the current $6.24 hourly wage).

The same calculations based on the other significant correlation are: b = .006, c = 4.377. Therefore,

$$Y = .006\ (X) = 4.377$$

For Omaha it becomes: .006 (347.38) + 4.377 or $6.46. The standard error is .839 or 84¢. Thus the estimate may be as low as $5.62 or as high as $7.30 68% of the time.

Since statisticians agree that regression analysis is of "limited value unless the variables in question are significantly correlated," this discussion will not report at this point these statistics for the other variables.**

Conclusions

Correlation analysis is used to measure the association of two variables. Regression analysis permits the prediction or estimation of one variable based on the measurement of a different variable. It permits the identification of deviant cases from a pattern of relationship--e.g. whether a

*For example, see G. David Garson, Handbook of Political Science Methods (Boston: Holbrook Press, Inc., 1971), p. 195.

**The data are provided, however, in the appendix.

city's bus drivers' hourly wages are different from what would be expected based upon an examination of another variable.

But the statistics of regression analysis are helpful only if the assumptions underlying them are valid and if the correlations between the variables are significant.

In the data presented to CAUR for analysis, only two variables are correlated significantly (at the .05 level) with bus driver hourly wages--proportion of workers using transit and the 1970 population. The regression analysis for the variable with the highest correlation indicates (a) the observed bus drivers' hourly wages are higher than the expected, and (b) the standard error of the estimate is large, especially in comparison to the differences in dispute before the Court of Industrial Relations. The second variable also has a large standard error of the estimate (although the estimate is higher than the actual hourly wage).

These limitations suggest that the use of these statistical concepts and data should be undertaken only with considerable caution.

APPENDIX

I. Correlation coefficients* between bus operator hourly wage and:

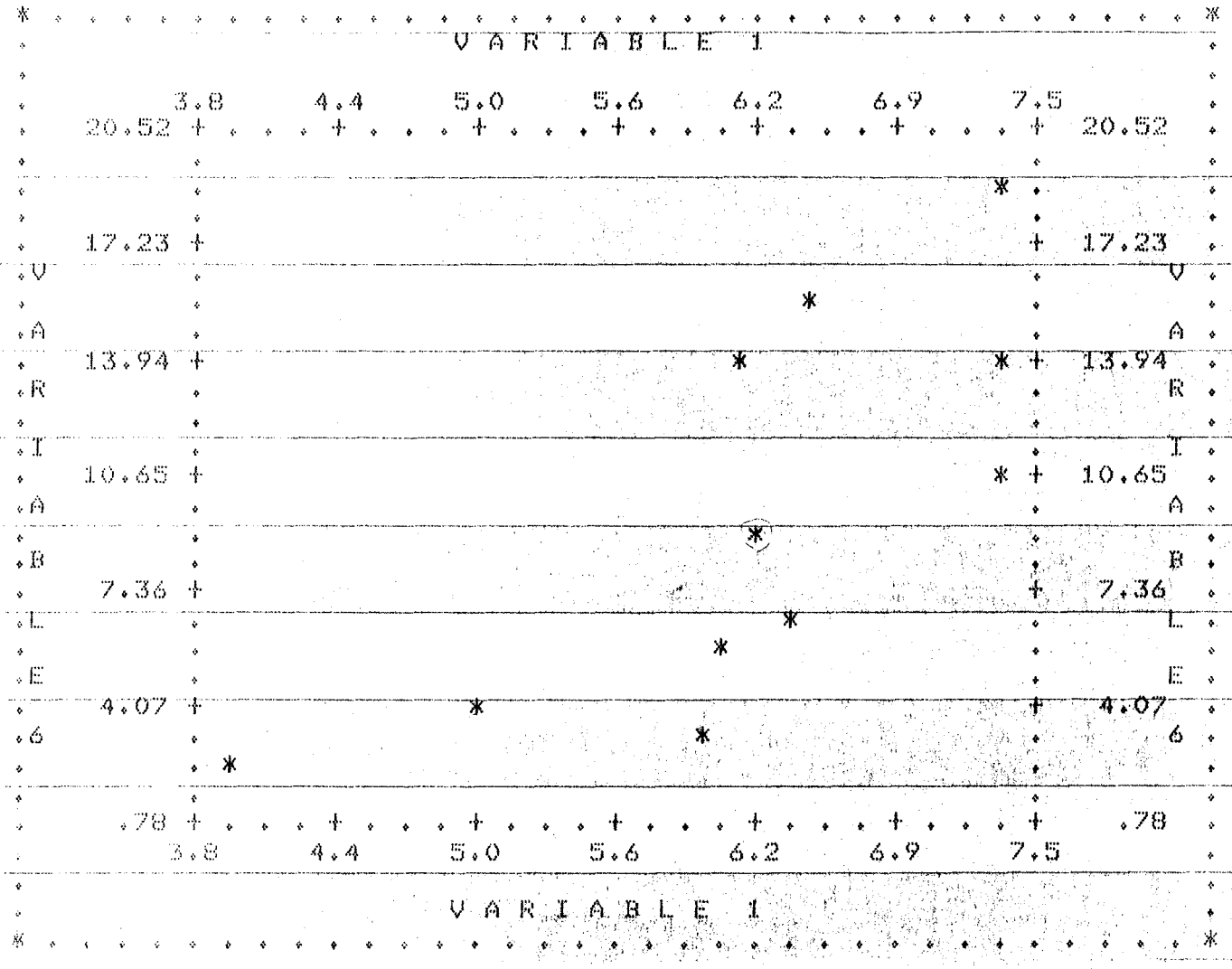
1.	Proportion workers using transit to work***	.747**
2.	1970 population	.617
3.	1973 population	.577
4.	July temperature	-.479
5.	January temperature	-.444
6.	Percent of unionization of plant workers	.430
7.	Population density	.389
8.	Per capita income	.386
9.	Median family income	.284
10.	Percent employment in manufacturing	.023
11.	Annual precipitation	.018

*For the 11 cities of : Omaha, Kansas City, Cincinnati, Minneapolis, Toledo, St. Paul, Akron, Wichita, Dayton, Des Moines, Lincoln

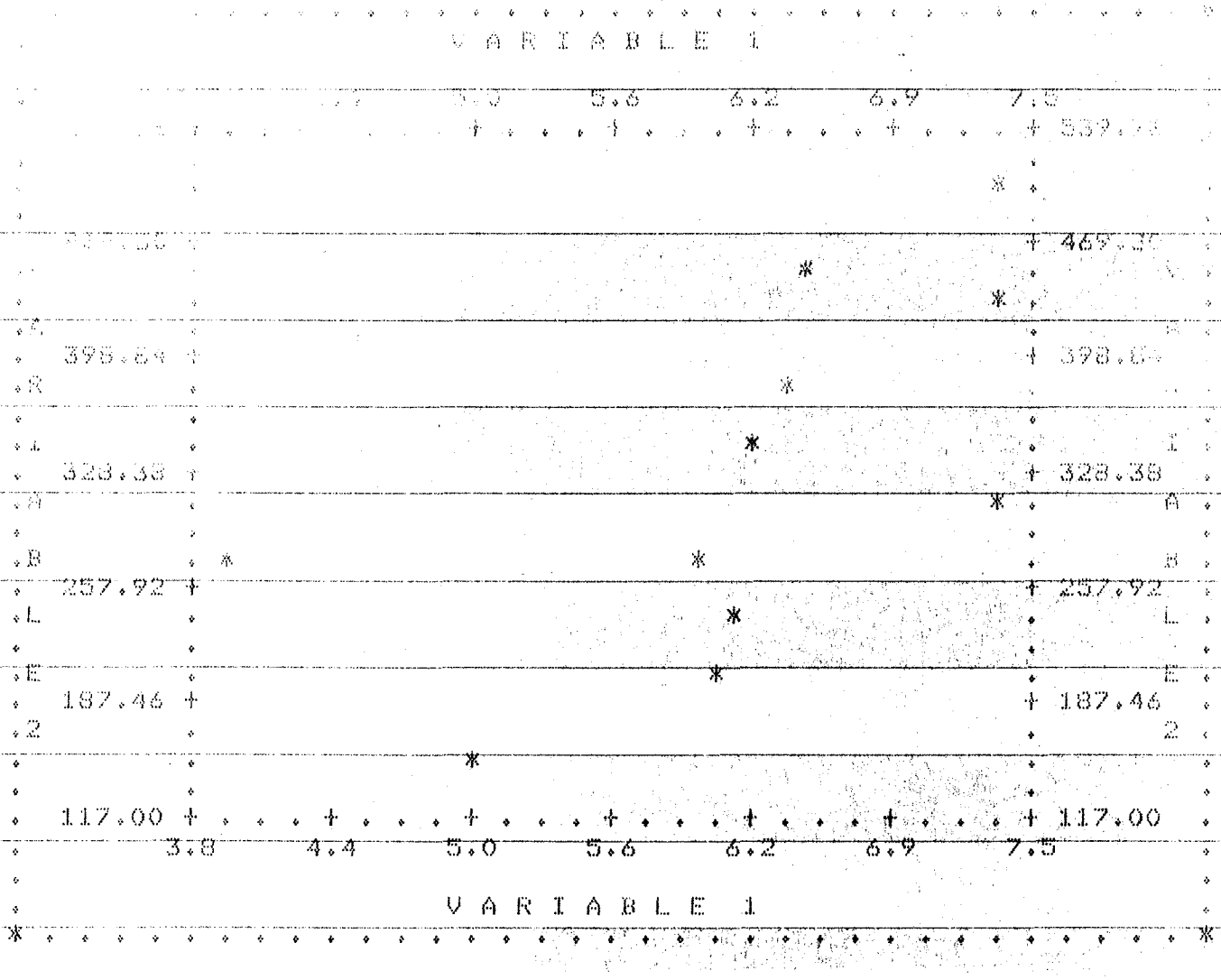
**Significant at the .05 level.

***Calculations based on 1970 population reported by Bureau of the Census (347,380) not that cited in the data provided (247,380 on p. 8).

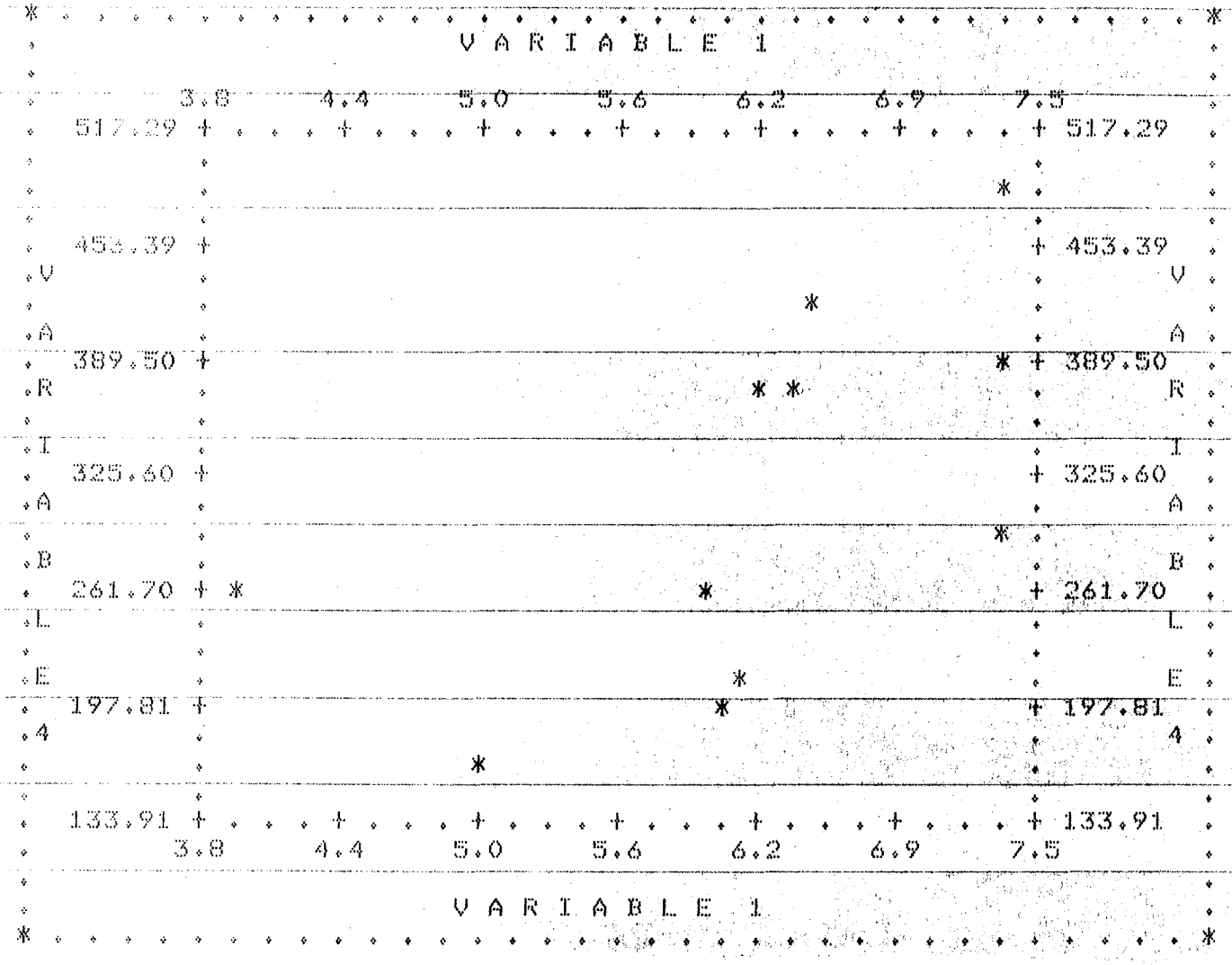
II. A. Scatter Diagram for Data on Bus Operator Hourly Wage and Proportion of Workers Using Transit to Work



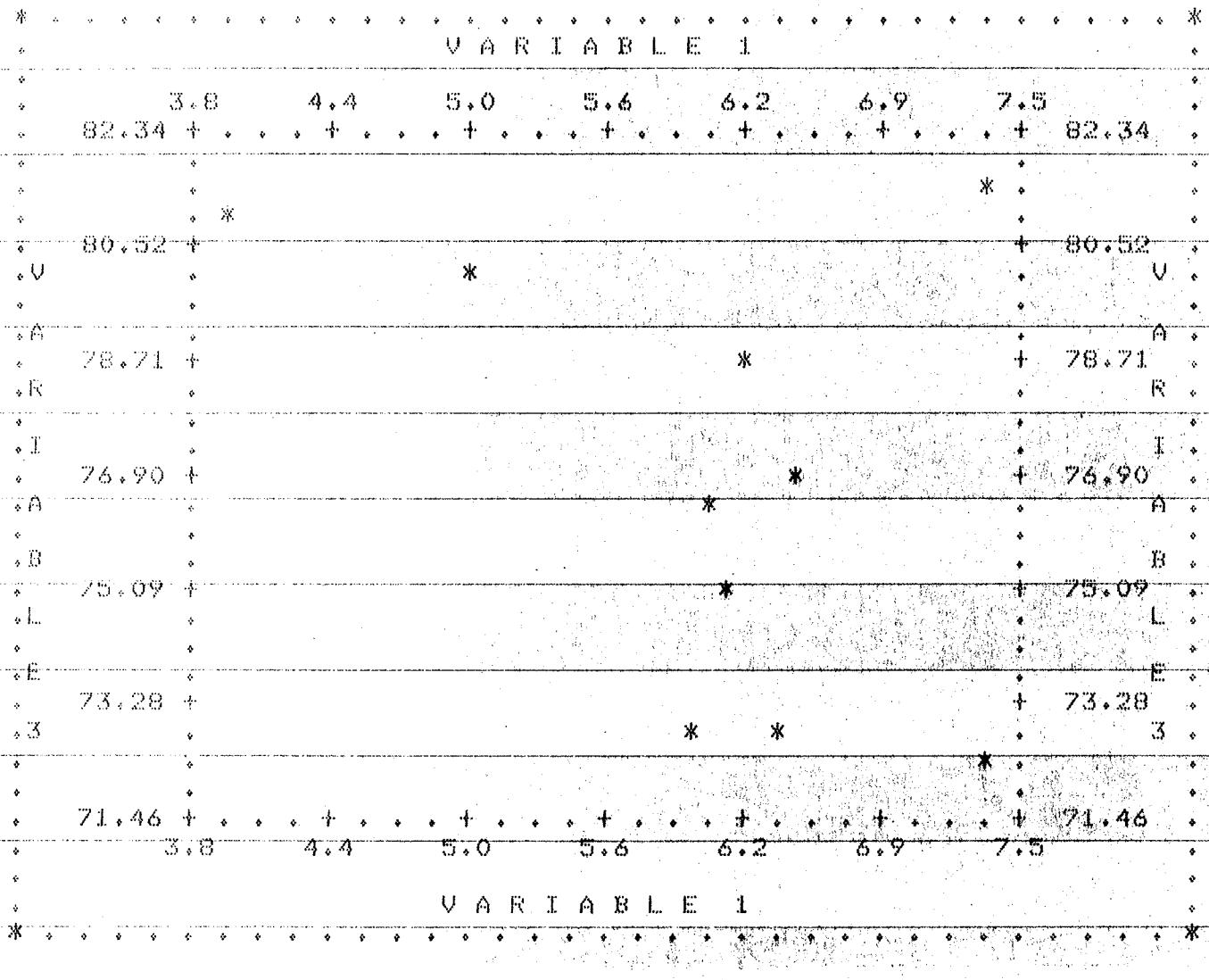
II. B. Scatter Diagram for Data on Bus Operator Hourly Wage and 1970 Population



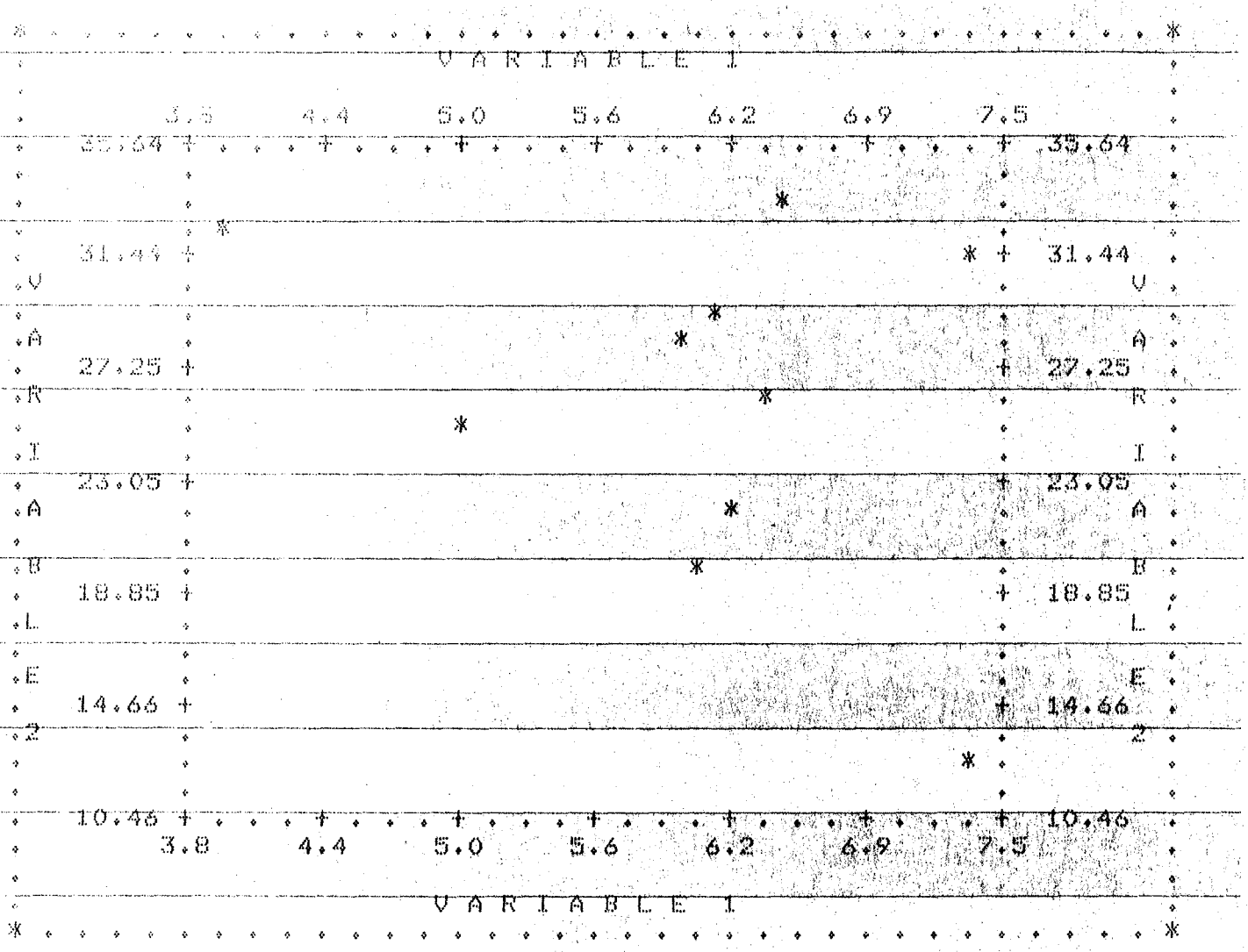
II. C. Scatter Diagram for Data on Bus Operator Hourly Wage and 1973 Population



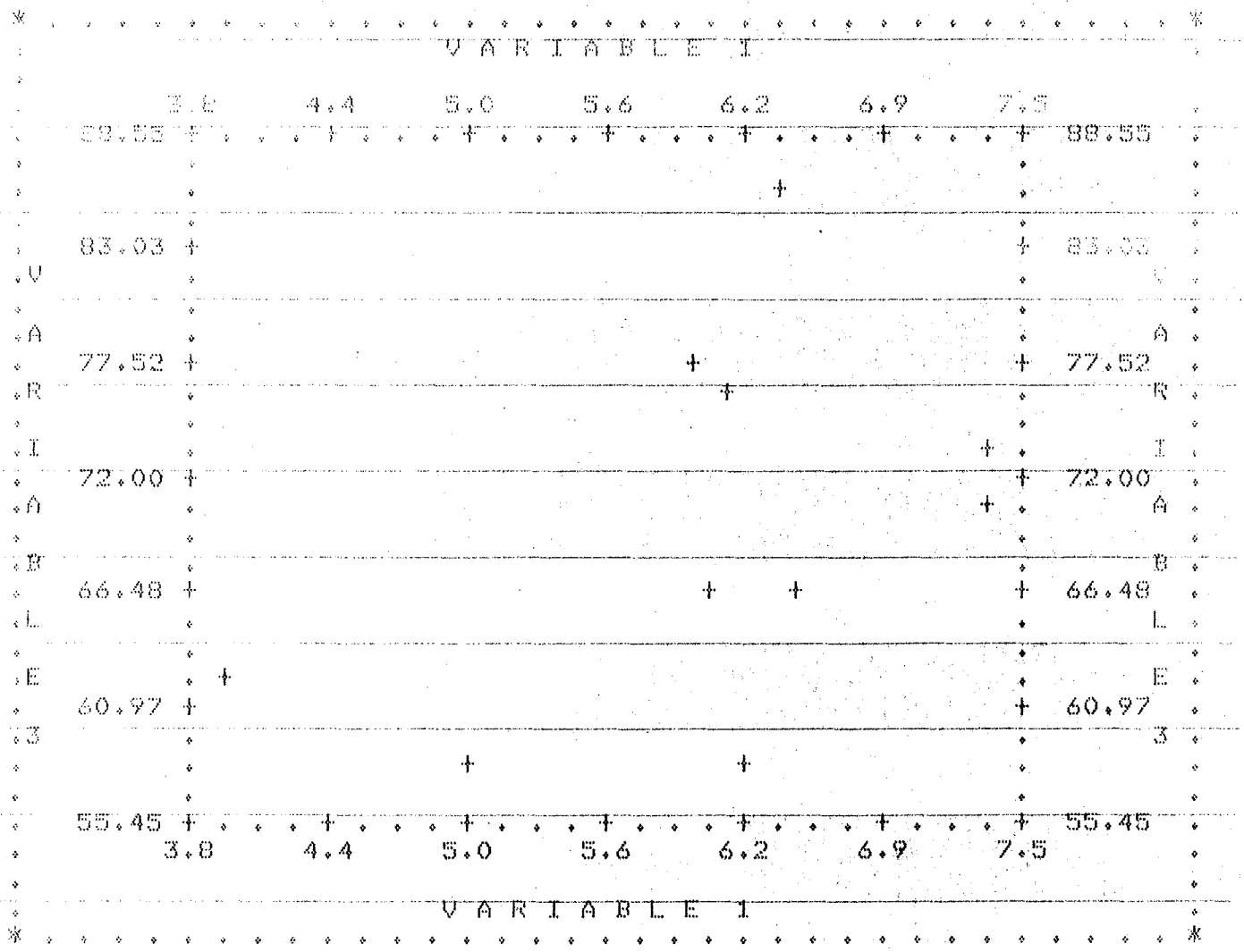
II. D. Scatter Diagram for Data on Bus Operator Hourly Wage and July Temperature



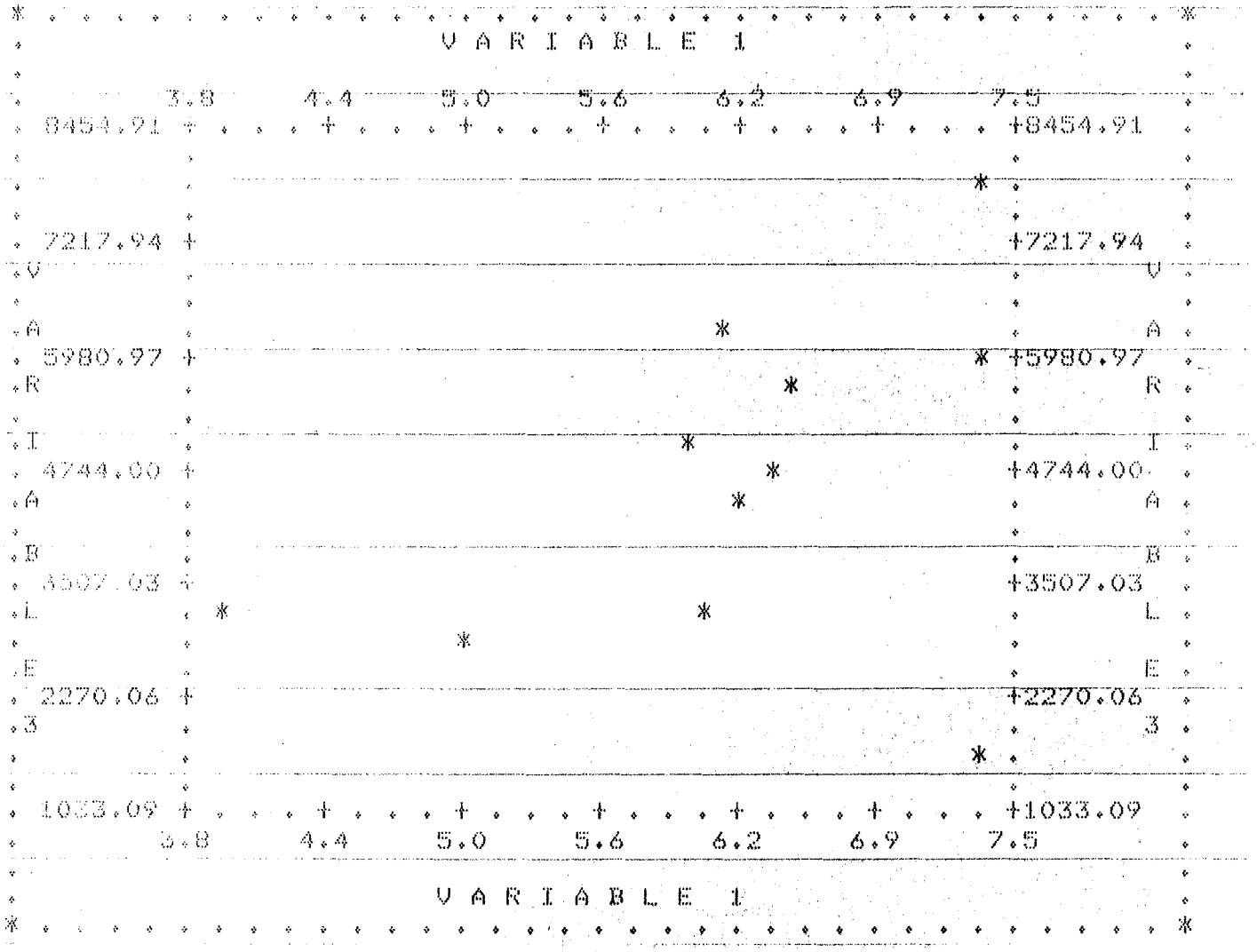
II. E. Scatter Diagram for Data on Bus Operator Hourly Wage and January Temperature



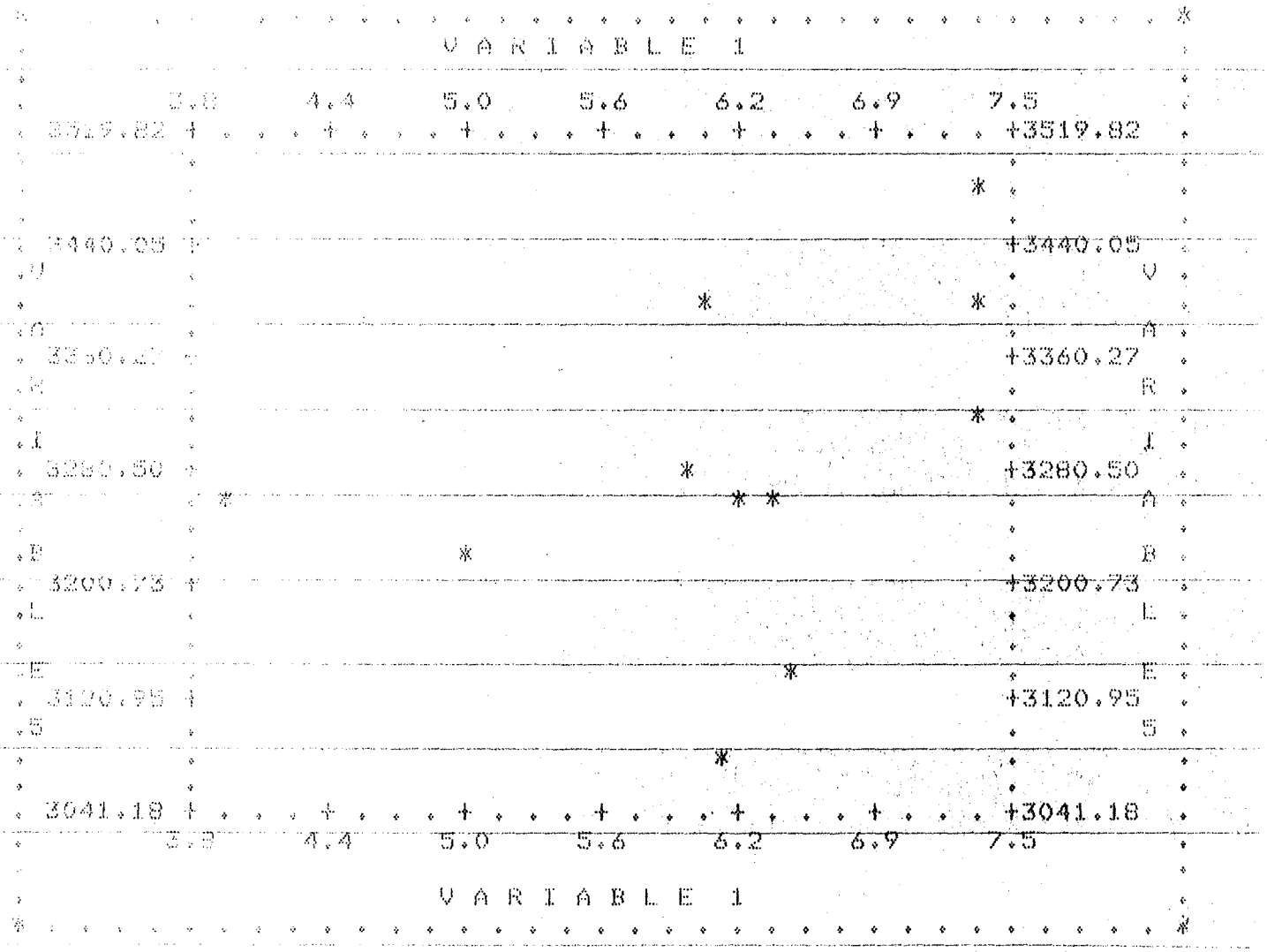
II. F. Scatter Diagram for Data on Bus Operator Hourly Wage and Percent of Unionization of Plant Workers



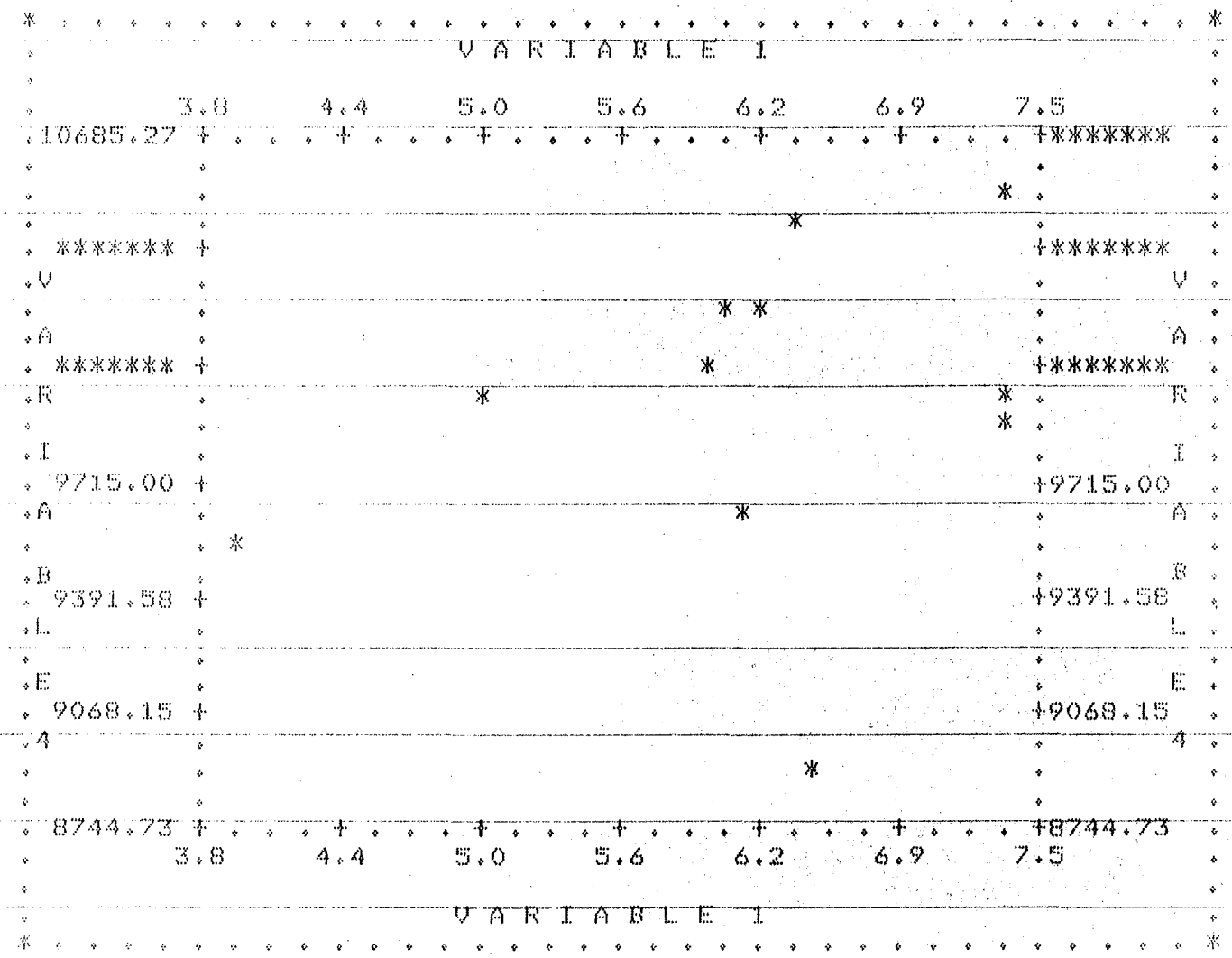
II. G. Scatter Diagram for Data on Bus Operator Hourly Wage and Population Density



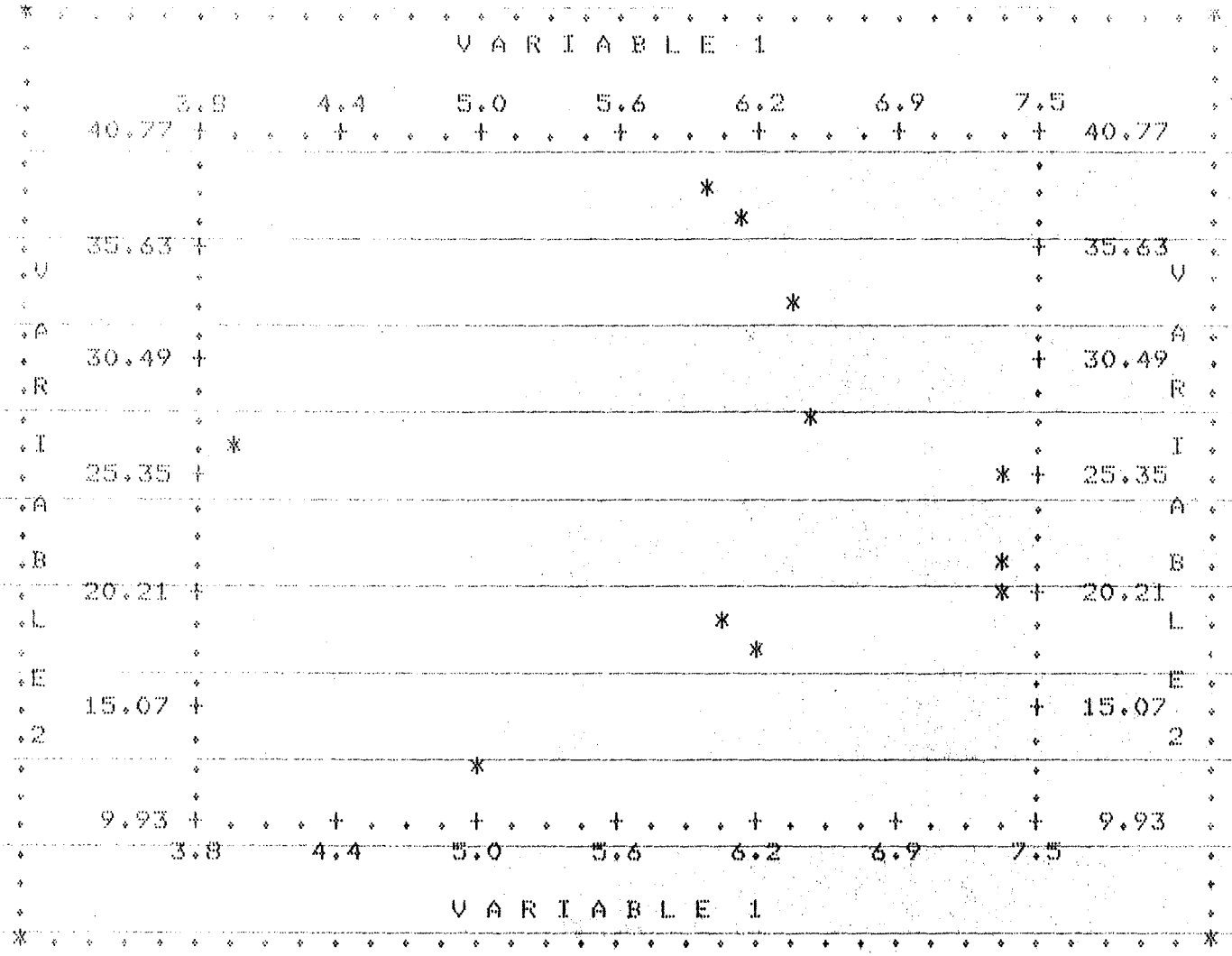
II. H. Scatter Diagram for Data on Bus Operator Hourly Wage and Per Capita Income



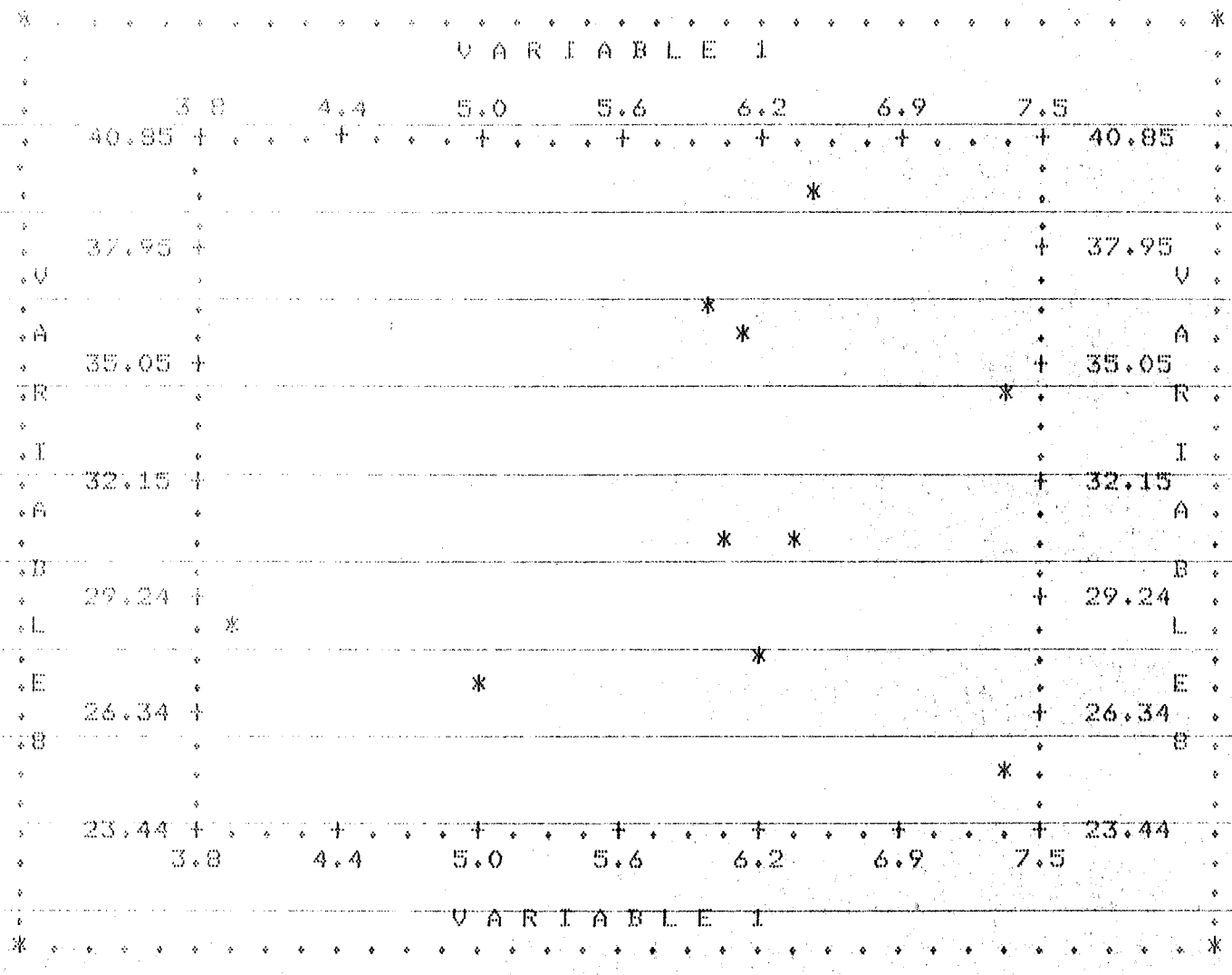
II. I. Scatter Diagram for Data on Bus Operator Hourly Wage and Median Family Income



II. J. Scatter Diagram for Data on Bus Operator Hourly Wage and Percent Employment in Manufacturing



II. K. Scatter Diagram for Data on Bus Operator Hourly Wage and Annual Precipitation



III. Regression coefficient, intercept, standard error of the estimate, and prediction of Omaha's bus operator hourly wage *

	Variable	Omaha Value (X)	Reg. Coef (b)	Inter-cept(a)	Predicted Bus Operator Hourly Wage (Y)**	Standard Error of the Estimate +
1.	% workers using transit++	9.3	.137	4.904	$6.18	.709
2.	1970 population++	347.38[a]	.006	4.377	6.46	.839
3.	1973 population	377.0	.006	4.448	6.71	.871
4.	July temperature	78.5	-.136	16.563	5.89	.936
5.	January temperature	22.3	-.060	7.711	6.37	.955
6.	% of unionization of plant workers	58.0	.051	2.617	5.58	.963
7.	Population density	4,529	.000	5.192	5.19	.982
8.	Per capita income	3,269	.003	-4.605	5.20	.984
9.	Median family income	10,206	.0006[b]	.093	6.22	1.022
10.	Percent employment in manufacturing	17.2	.003	6.137	6.19	1.066
11.	Annual precipitation	27.56	.004	6.097	6.21	1.066

*Observed Omaha bus operator hourly wage was $6.24.

**$Y = b(X) + c$

+The standard error provides 68% confidence limits, and two standard errors provide 95% confidence limits.

++Correlation significant at .05 level.

[a]All calculations based on corrected data; Omaha's 1970 population was 347,380 not the 247,380 reported on p. 8.

[b]The regression coefficient is .000616; if rounded to only three places (.001) an estimated hourly wage of $10.30 results; if carried to all six places the estimate is $6.29.