tion required (e.g. moments, estimation, application) can be cross-classified, as they are common to all distributions.

REFERENCES

1. Box, G. E. P. (1949). *Biometrika,* **36,** 317–346.
2. Box, G. E. P. and Andersen, S. L. (1955). *J. Roy Statist. Soc. Ser. B,* **17,** 1–26.
3. Bowman, K. O. and Shenton, L. R. (1965, 1966). Reports K–1633, K–1643, ORNL–4005, Union Carbide Corporation.
4. Haight, F. A. (1967). *Handbook of the Poisson Distribution,* New York: John Wiley and Sons.
5. James, A. T. (1954, 1960, 1964). *Ann. Math Statist.,* **25,** 40–75; **31,** 151–8; **35,** 475–97.
6. Johnson, N. L. and Kotz, S. (1969, 1970, 1972). *Distributions in Statistics,* Vol. I (Discrete). Vols. II and III (Continuous Univariate), Vol. IV (Continuous Multivariate), New York: John Wiley and Sons.
7. Kotz, S. and Johnson, N. L. (1969). *Distribution Theory in Statistical Literature,* Proc. 37th Session of the ISI, 303–305.
8. Lancaster, H. O. (1969). *The Chi-squared Distribution,* New York: John Wiley and Sons.
9. Milton, R. C. (1969). Computer Implementation of Distribution Function Algorithms. *Proc. Conference on Statistics and Computers,* University of Wisconsin, Madison, pp. 181–198.
10. Sichel, H. S. (1947). *J. Roy. Statist. Soc. Ser. A,* **110,** 337–47; (1949) *Biometrika,* **36,** 404–25.
11. Weiss, L. and Wolfowitz, J. (1966, 1968). *Teoriya Veroyatnostei i ee Primeneniya,* **11,** 68–93; **13,** 657–662. (English version of the journal: **11:** 58–81; **13,** 622–627.)

Graphs in Statistical Analysis*

1973!

F. J. ANSCOMBE**

Graphs are essential to good statistical analysis. Ordinary scatterplots and "triple" scatterplots are discussed in relation to regression analysis.

1. *Usefulness of graphs*

Most textbooks on statistical methods, and most statistical computer programs, pay too little attention to graphs. Few of us escape being indoctrinated with these notions:

(1) numerical calculations are exact, but graphs are rough;
(2) for any particular kind of statistical data there is just one set of calculations constituting a correct statistical analysis;
(3) performing intricate calculations is virtuous, whereas actually looking at the data is cheating.

A computer should make *both* calculations *and* graphs. Both sorts of output should be studied; each will contribute to understanding.

Graphs can have various purposes, such as: (i) to help us perceive and appreciate some broad features of the data, (ii) to let us look behind those broad features and see what else is there. Most kinds of statistical calculation rest on assumptions about the behavior of the data. Those assumptions may be false, and then the calculations may be misleading. We ought always to try to check whether the assumptions are reasonably correct; and if they are wrong we ought to be able to perceive in what ways they are wrong. Graphs are very valuable for these purposes.

Good statistical analysis is not a purely routine matter, and generally calls for more than one pass through the computer. The analysis should be sensitive both to peculiar features in the given numbers and also to whatever background information is available about the variables. The latter is particularly helpful in suggesting alternative ways of setting up the analysis.

Thought and ingenuity devoted to devising good graphs are likely to pay off. Many ideas can be gleaned from the literature, of which a sampling is listed at the end of this paper. In particular, Tukey [7, 8] has much to say on the topics presented here.

A few simple types of statistical analysis are now considered.

2. *Regression analysis—the simplest case*

Suppose we have values for one "dependent" variable y and one "independent" (exogenous, predictor) variable x. Before anything else is done, we should scatterplot the y values against the x values and *see* what sort of relation there is—if any. Many different kinds of things can happen:—

(1) the (x, y) points lie nearly on a straight line;
(2) the (x, y) points lie nearly on a smooth curve, not a straight line;
(3) the y-values are scattered, without relation to the x-values;
(4) something intermediate between (1) or (2) and (3);
(5) most of the (x, y) points lie close to a line or smooth curve, but a few are scattered a long way away.

Case (5) is particularly interesting, because there is an effect to be noticed, but the ordinary calculations for linear regression may miss it. Whenever we see "outliers", it is usually wise first to check that the

* Prepared in connection with research supported by the Army, Navy, Air Force and NASA under a contract administered by the Office of Naval Research.

** Dept. of Statistics, Yale Univ., Box 2179, Yale Station, New Haven, Conn. 06520.

values used really are correct, that is, not copied wrongly nor obviously faulty in some way. Then, if we are satisfied that these readings are authentic, we may perhaps set them aside for special study, and fit a regression relation to the remainder of the data. Special study of the outliers may prove very rewarding.

Case (1) would usually be considered ideal. Case (2) can sometimes be brought back to case (1) by transforming the x-scale or the y-scale or both.

The ordinary least-squares regression calculation is based on the following theoretical description or "model": the given number pairs (x_i, y_i) are related by

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \qquad (i = 1, 2, \ldots, n), \quad \text{(A)}$$

where β_0 and β_1 are constants and the "errors" $\{\epsilon_i\}$ are drawn independently from a "normal" (Gauss-Laplace) probability distribution having zero mean and constant variance. The regression calculation leads to estimates b_0 and b_1 for β_0 and β_1, to the "fitted values"

$$\hat{y}_i = b_0 + b_1 x_i,$$

and to the "residuals"

$$e_i = y_i - \hat{y}_i.$$

The sum of squares of the latter, generally called the "residual sum of squares" or "error sum of squares", leads to an estimate of the variance of the distribution of errors. If the theoretical description were exactly correct (and all calculation were exact, without round-off error), these calculations would be entirely satisfactory, in the sense that b_0, b_1 and the residual sum of squares, together with the number of readings n and the first two moments of the x-values, would constitute sufficient statistics for the unknowns and could substitute for the original data for all purposes with no loss of information. In practice, we do not know that the theoretical description is correct, we should generally suspect that it is not, and we cannot therefore heave a sigh of relief when the regression calculation has been made, knowing that statistical justice has been done.

After the regression calculation, the residuals $\{e_i\}$ should be plotted against the $\{x_i\}$. One might think this would show nothing that could not be seen in the original plot of $\{y_i\}$ against $\{x_i\}$. However, the residual plot will probably have a larger scale for the ordinates, and with the linear regression removed the residual behavior is easier to see. Usually it is a good idea to specify that the residual plot should be of the residuals $\{e_i\}$ against the fitted values $\{\hat{y}_i\}$, rather than $\{x_i\}$, with the *same scale* for ordinates and abscissas. This plot, besides showing how the residuals behave in relation to the x-values, also from its overall shape shows at a glance the relative dispersion of fitted values and residuals. In the decomposition

$$y_i = \hat{y}_i + e_i$$

$$\text{(observation} = \text{fitted value} + \text{residual)},$$

hopefully the fitted values follow the observations closely and have a greater variability than the residuals. One should be aware of their relative contributions.

If the theoretical description of the observations were exactly true, the residuals would appear to be normally distributed with zero mean and common variance, the same for all x-values. [That statement is not quite correct, but near enough for most practical purposes. The residuals would usually not have exactly equal variances, and they would be variously correlated.] Things to look for in a plot of $\{e_i\}$ against $\{\hat{y}_i\}$ or $\{x_i\}$ are:—

(1) a few of the residuals much larger in magnitude than all the others—outliers;
(2) a curved regression of residuals on fitted values;
(3) progressive change in the variability of the residuals as the fitted values increase;
(4) a skew (or other nonnormal) distribution of the residuals.

Sometimes, if we are lucky, effects (2), (3) and (4) can be removed simultaneously by a transformation of the scale in which y is expressed, as by taking logarithms. Alternatively, effect (2) may be allowed for by transforming the x-scale, or by adding another term on the right side of the theoretical description (A), making it perhaps

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i.$$

Instead of looking at a scatterplot of $\{e_i\}$ against $\{\hat{y}_i\}$, we could detect effects such as those just listed by calculating suitable test statistics, and we could assess their significance. But the plot shows a variety of features quickly and vividly, and formal tests often seem unnecessary.

There is indeed another reason for examining a scatterplot of residuals against fitted values, that may be important even when there is no indication of inadequacy in the theoretical description (A). Possibly one (or a few) observations have x-values widely separated from the others, leading to (one or more) outliers among the fitted values. Even though (A) should seem to fit all observations satisfactorily, with no outliers among the residuals, we may feel less comfortable about postulating (A) and basing conclusions on it, than if there had been no greatly outlying fitted value. That is because an outlying x-value contributes much more to the determination of the regression coefficient than other x-values. If an observation with an outlying x-value were affected by some special circumstance, not common to other observations, our fitted regression relation might be misleading. Often the y-value corresponding to an outlying x-value could be altered considerably without much effect on the goodness of fit of the regression relation but with marked effect on the estimated relation itself. We are usually happier about asserting a regression relation if the relation is still apparent after a few observations (any ones) have been deleted—that is, we are happier if the regression relation seems to permeate all the observations and does not derive largely from one or two.

All these various features that can so greatly change the significance we attach to a calculated regression are

invisible if we see only the usual quadratic summaries—the regression line, the analysis of variance, the multiple correlation coefficient R^2.

3. *An example*

Some of these points are illustrated by four fictitious data sets, each consisting of eleven (x, y) pairs, shown in the table. For the first three data sets the x-values are the same, and they are listed only once.

Data set		1-3	1	2	3		4	4
Variable		x	y	y	y		x	y
Obs. no.	1	10.0	8.04	9.14	7.46		8.0	6.58
	2	8.0	6.95	8.14	6.77		8.0	5.76
	3	13.0	7.58	8.74	12.74		8.0	7.71
	4	9.0	8.81	8.77	7.11		8.0	8.84
	5	11.0	8.33	9.26	7.81		8.0	8.47
	6	14.0	9.96	8.10	8.84		8.0	7.04
	7	6.0	7.24	6.13	6.08		8.0	5.25
	8	4.0	4.26	3.10	5.39		19.0	12.50
	9	12.0	10.84	9.13	8.15		8.0	5.56
	10	7.0	4.82	7.26	6.42		8.0	7.91
	11	5.0	5.68	4.74	5.73		8.0	6.89

TABLE. Four data sets, each comprising 11 (x, y) pairs.

Each of the four data sets yields the same standard output from a typical regression program, namely

Number of observations $(n) = 11$
Mean of the x's $(\bar{x}) = 9.0$
Mean of the y's $(\bar{y}) = 7.5$
Regression coefficient (b_1) of y on $x = 0.5$
Equation of regression line: $y = 3 + 0.5\,x$
Sum of squares of $x - \bar{x} = 110.0$
Regression sum of squares $= 27.50$ (1 d.f.)
Residual sum of squares of $y = 13.75$ (9 d.f.)
Estimated standard error of $b_1 = 0.118$
Multiple $R^2 = 0.667$

These calculations express in various (redundant) ways the sufficient statistics for the theoretical description (A), when that is assumed to be correct. Some typical computer programs also yield a print-out of the residuals, in the order in which the data were entered. Since in the present case the data have been listed in a random order, probably little would be seen if the eye were run down such a print-out (especially if it were in abominable floating-point notation).

The data sets are graphed in the figures, together with the fitted line. Figure 1, corresponding to data set 1, is the kind of thing most people would see in their mind's eye, if they were presented with the above calculated summary. The theoretical description (A) seems to be perfectly appropriate here, and the calculated summary seems fair and adequate. Figure 2 suggests forcefully that data set 2 does not conform with the theoretical description (A), but rather y has a smooth curved relation with x, possibly quadratic, and there is little residual variability. Figure 3 similarly suggests that (A) is not a good description for data set 3: all but one of the observations lie close to a straight line (not the one yielded by the standard regression calculation), namely

$$y = 4 + 0.346\,x;$$

and one observation is far from this line. Those are the essential facts that need to be understood and reported.

Figure 4, like Figure 1, shows data apparently conforming well with the theoretical description (A). If all observations are considered genuine and reliable, data set 4 is just as informative about the regression relation as data set 1; there is no reason to prefer either to the other. Yet in most circumstances we should feel that there was something unsatisfactory about data set 4. All the information about the slope of the regression line resides in one observation—if that observation were deleted the slope could not be estimated. In most circumstances we are not quite sure that every observation is reliable. If any one observation were discredited and therefore deleted from data set 1, the remainder would tell much the same story. That is not so for data set 4. Thus the standard regression calculation ought to be accompanied by a warning that one observation has played a critical role.

Each of data sets 2, 3, 4 illustrates a peculiar effect in an extreme form. In less extreme forms such effects are often encountered in statistical analysis. For an example of the last effect: in a study (to be published

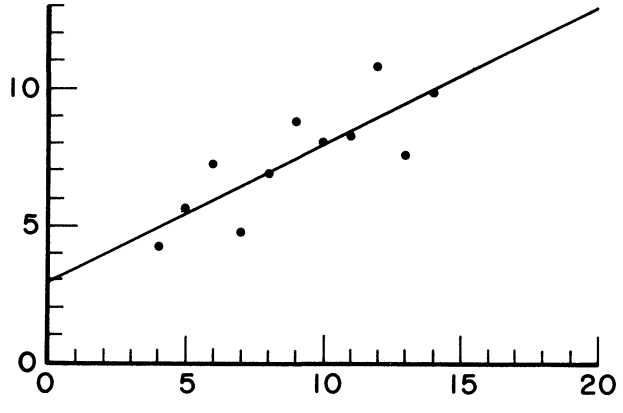

Figure 1

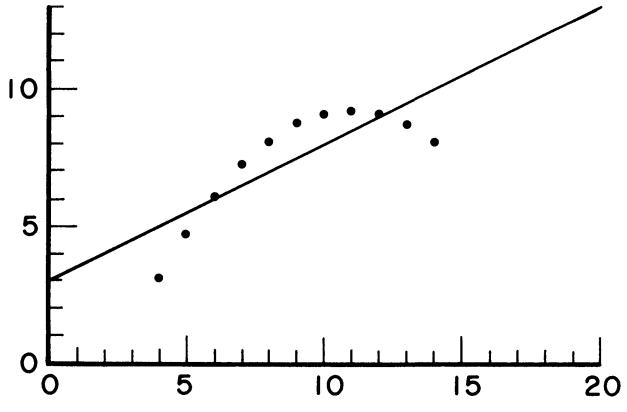

Figure 2

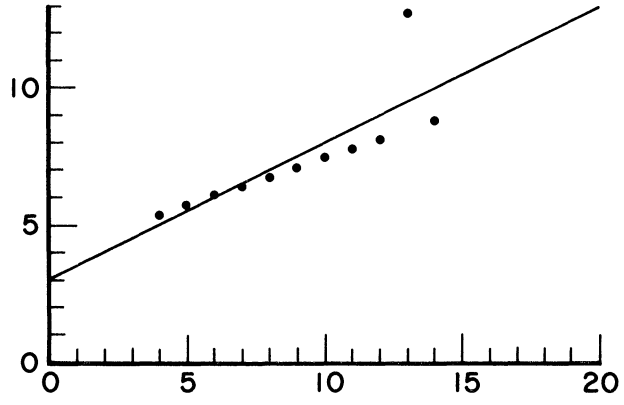

Figure 3

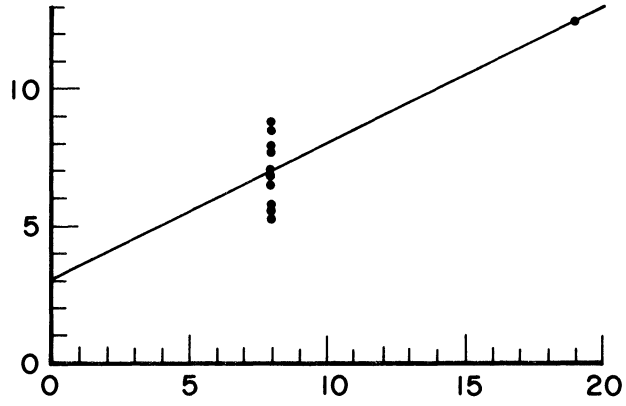

Figure 4

elsewhere) of per capita expenditures on public school education in each of the fifty states of the Union, together with the District of Columbia, it was found that the expenditures had a satisfactory linear regression on three likely predictor variables, with multiple R^2 about 0.7 and well behaved residuals. However, one of the states, namely Alaska, was seen to have values for the predictor variables rather far removed from those of the other states, and therefore Alaska contributed rather heavily to determining the regression relation. Of course Alaska is an abnormal state, and the thought immediately occurs that perhaps Alaska should be excluded from the study. But there are other extraordinary states, Hawaii, the District of Columbia (counted here as a state), California, Florida, New York, North Dakota, . . . Where does one stop? Rather than merely exclude Alaska, a preferable course seems to be to report the regression relation when all states are included, but add that Alaska has contributed heavily and say what happens if Alaska is omitted—the regression relation is not greatly changed, but the standard errors are increased somewhat and multiple R^2 is reduced below 0.6. We need to understand *both* the regression relation visible in all the data *and also* Alaska's special contribution to that relation.

4. *More general regression analysis*

Much of what has been said about regression of one dependent on one independent variable applies to more general regression analyses. Suppose there is one dependent variable y but two "independent" variables $x^{(1)}$ and $x^{(2)}$, so that the theoretical description reads

$$y_i = \beta_0 + \beta_1 x_i^{(1)} + \beta_2 x_i^{(2)} + \epsilon_i, \qquad \text{(B)}$$

where the β's are constants and the ϵ's are distributed as before.

We cannot simply make on a two-dimensional surface a three-dimensional plot of y against $x^{(1)}$ and $x^{(2)}$ simultaneously. There are indeed expensive visual devices for suggesting such a thing. If we confine ourselves to what can be done with a line printer or typewriter terminal, there are two approaches to visualizing relations between the three variables y, $x^{(1)}$ and $x^{(2)}$, before any regression calculation.

(a) Make ordinary scatterplots of the three pairs of variables, y against $x^{(1)}$, y against $x^{(2)}$, $x^{(1)}$ against $x^{(2)}$. The third of these shows whether it will be possible to distinguish the effects of $x^{(1)}$ and $x^{(2)}$ on y. For if $x^{(1)}$ and $x^{(2)}$ are closely related to each other, one of them having a regression (not necessarily linear) on the other with little residual variation, then any apparent relation of y with $x^{(1)}$ and $x^{(2)}$ may perhaps be expressible equally well as a relation of y with either $x^{(1)}$ or $x^{(2)}$ alone.

(b) Make a scatterplot of two of the variables, say $x^{(1)}$ and $x^{(2)}$, marking each point by a symbol that roughly indicates the value of the third variable y. The values of the third variable can be coded numerically, say by dividing the range into ten intervals and representing values, according to the interval they fall in, by single digits, 0, 1, 2, . . . , 9—or possibly by dividing the range into not more than twenty-six intervals and using letters of the alphabet. Alternatively, values of the third variable may be coded by symbols whose physical appearance (size and blackness) indicates magnitude—for example, with an APL typeball, symbols of increasing weight such as

. ∘ ○ ⊖ □ ⊟

or these representing steps from large-negative to large-positive (M standing for minus and P for plus)

M M - ∘ + P P

Such a plot is equivalent to an ordinary scatterplot of the first two variables and also indicates, well enough for many purposes, how the third variable is related to the other two. This kind of plot will be called a *triple scatterplot* (TSCP).

After the ordinary regression calculations have been made, yielding the regression coefficients b_0, b_1, b_2, the fitted values and the residuals, the single most useful plot is an ordinary scatterplot of residuals against fitted values, preferably on the same scale. Interpretation is as indicated before.

Another possibility is to make a TSCP, taking as the first two variables the contributions of $x^{(1)}$ and $x^{(2)}$ to the fitted values, that is, plot $\{b_1 x_i^{(1)}\}$ against $\{b_2 x_i^{(2)}\}$ on the same scale, with the residuals $\{e_i\}$ coded as the third variable. This plot can show association of residual behavior with $x^{(1)}$ and $x^{(2)}$ individually.

To study the dependence of y on one of the independent variables, say $x^{(1)}$, with the effect of the other eliminated, one may scatterplot $\{y_i - b_2 x_i^{(2)}\}$ against $\{x_i^{(1)}\}$. This would be useful in planning a transformation of the $x^{(1)}$-scale.

When we pass from a regression problem with only two "independent" variables to one with many, we find it harder to see all that is going on by looking at graphs. But that is as it should be—the possibilities are now so much greater. The likelihood that we fool ourselves by *only* carrying out some ordinary regression calculations is much greater too. Usually when there are many "independent" variables they are mutually related and we are interested in performing regression on subsets of them, possibly by a "stepwise" procedure; so even the standard calculation is not so simple.

In any case, whenever a regression calculation has been carried out, whether on all the "independent" variables or on a subset of them, it will be useful to see a simple scatterplot of residuals against fitted values (on the same scale).

If the independent variables are separated into two sets, we may be interested to see a TSCP, in which the two coordinates represent the contributions of each of the two sets to the fitted values (on the same scale) and the plotting code represents the residuals.

5. *Two-way tables*

The analysis of a two-way table by calculating row means, column means, residuals and what R. A. Fisher called the analysis of variance, may be regarded as a special instance of regression analysis. The structure is now sufficiently rich that graphical presentation in advance of numerical calculation is probably not too useful. But after the calculations the same sorts of graphical treatment as for ordinary regression have the same effectiveness. Residuals may be scatterplotted against fitted values on the same scale. Row effects can be plotted against column effects, on the same scale, in a TSCP with coded residuals. (It was Tukey's elegant use of a kind of TSCP for two-way tables that introduced me to the idea; see Chapter 16 in [7].) If the rows or columns have a meaningful natural order, the residuals should also be presented in that order.

Rectangular tables (crossclassifications) in two or more dimensions, with some modes of classification perhaps "nested" rather than "crossed", are of common occurrence. Whenever any set of main effects and interactions has been calculated, the residuals should be scatterplotted against the fitted values, and various sorts of TSCP may be interesting.

This article is emphatically not a catalog of useful graphical procedures in statistics. Its purpose is merely to suggest that graphical procedures are useful. Only two types of graph have been mentioned, the ordinary scatterplot and the triple scatterplot, and these have been considered in only one sort of context (regression). There are other types of graphs and display devices that can make quantitative relations visible and comprehensible, and other sorts of statistical tasks than regression.

6. *Implementation*

Graphical output such as described above is readily available to anyone who does his own programming. I myself habitually generate such plots at an APL terminal, and have come to appreciate their importance. A skilled Fortran or PL/1 programmer, with an organized library of subroutines, can do the same (on a larger scale).

Unfortunately, most persons who have recourse to a computer for statistical analysis of data are not much interested either in computer programming or in statistical method, being primarily concerned with their own proper business. Hence the common use of library programs and various statistical packages. Most of these originated in the pre-visual era. The user is not showered with graphical displays. He can get them only with trouble, cunning and a fighting spirit. It's time that was changed.

REFERENCES

[1] Anderson, Edgar, "A semigraphical method for the analysis of complex problems," *Proceedings of the National Academy of Sciences*, 13 (1957), 923–927. Reprinted in *Technometrics*, 2 (1960), 387–391.

[2] Andrews, D. F., "Plots of high-dimensional data," *Biometrics*, 28 (1972), 125–136.

[3] Bachi, Roberto, *Graphical Rational Patterns*, Jerusalem: Israel Universities Press, 1968.

[4] Bertin, Jacques, *Sémiologie Graphique*, Paris: Mouton and Gauthier-Villars, 1967.

[5] Daniel, Cuthbert, "Use of half-normal plots in interpreting factorial two-level experiments," *Technometrics*, 1 (1959), 311–341.

[6] Draper, N. R., and Smith, H., *Applied Regression Analysis*, New York: Wiley, 1966.

[7] Tukey, John W., *Exploratory Data Analysis*, limited preliminary edition, three volumes, Reading: Addison-Wesley, 1970–71.

[8] Tukey, John W., "Some graphic and semigraphic displays," *Statistical Papers in Honor of George W. Snedecor* (ed. T. A. Bancroft), Ames: Iowa State University Press, 1972, pp. 293–316.

[9] Tukey, John W., and Wilk, M. B., "Data analysis and statistics: techniques and approaches," *The Quantitative Analysis of Social Problems* (ed. E. R. Tufte), Reading: Addison-Wesley, 1970, pp. 370–390.

[10] Wilk, M. B., and Gnanadesikan, R., "Probability plotting methods for the analysis of data," *Biometrika*, 55 (1968), 1–17.

[11] "Statistical analysis, special problems of, I. Outliers, II. Transformations of data," *International Encyclopedia of the Social Sciences*, Macmillan and Free Press, 1968, vol. 15, pp. 178–193.