MODELS FOR PREDICTION PURPOSES

By

A. E. MAXWELL

Maudsley Hospital, London, S.E.5

IT is natural for psychiatrists and others dealing with mental patients to enquire what the prognosis is for a patient showing certain symptoms or syndromes of symptoms—taking age and the history of the patient's illness into account. The very fact that the symptoms generally form syndromes points to their interdependence to a greater or lesser degree, but this interdependence may not be a straightforward linear relationship such as is revealed by an ordinary correlation coefficient for the symptoms may interact with each other in a complicated way. When this is the case the problem of prediction is difficult from a statistical as well as from a psychiatric point of view. In this paper an effort is made to clarify the statistical problem by describing briefly and by comparing two statistical models which may be used for predictive purposes. The models are introduced by means of an example using data for patients who have attended this hospital. For reasons of clarity the number of variables included in the study is a bare minimum.

THE CORRELATION MODEL

For a sample of 819 psychiatric patients, described in some detail in an earlier article in this journal (Trouton and Maxwell, 1956), the presence or absence of the three symptoms:

d depressed
a anxious
g delusions of guilt

were noted. Two-by-two contingency tables for each pair of symptoms were drawn up and the tetrachoric correlations between the symptoms obtained. These were:

	d	*a*	*g*
d	—	0·385	0·600
a	0·385	—	0·289
g	0·600	0·289	—

The tetrachoric correlations between each symptom and the criterion, "recovered or much improved at time of discharge", which we will denote by the letter "c", were also found. They are:

	d	*a*	*g*
c	0·318	0·206	0·278

In an effort to assess the relative effects of the symptoms as regards recovery the customary approach is to employ a multiple regression model. Using it regression weights, known as partial regression coefficients, are calculated. These are the optimal weights to apply to the independent variables (in our case the symptoms) to obtain the highest possible correlation—the multiple correlation—between the independent variables and the criterion. These

https://doi.org/10.1192/bjp.105.439.463 Published online by Cambridge University Press

regression weights then give us a measure of the relative importance of the independent variables for predicting the criterion. For our data the regression coefficients with their standard errors are:

Variable	Regression Weight	Statistical Error	Significance Level
d	0·206	0·04276	P<·001
a	0·089	0·03572	P<·02
g	0·128	0·04122	P<·01

The method by which they are obtained is given in most textbooks on general statistics.

Each of the regression weights is seen to be significant and the order of their importance for predicting the patients' recovery is, 1st—depression, 2nd—delusions of guilt, and 3rd—anxious. If then a patient is *depressed* his prognosis is slightly better than if he is *anxious*, etc. But what does the model tell us about the odds in his favour? The multiple regression coefficient throws some light on this problem. Each of the three symptoms under discussion is correlated only to a small degree with the criterion—the largest of the three correlations is 0·318. However, the multiple correlation between the three combined when the regression, or best weights, are used is 0·346, which is somewhat larger than 0·318. If the multiple correlation coefficient is squared and multiplied by 100 the figure 11·9 is obtained, and this is the percentage of the variance of the criterion which we can predict from a knowledge of the individual symptoms. This immediately tells us that any prediction we make about a patient's recovery from the presence or absence of these three symptoms only is unlikely to be very accurate. To improve our prediction we would require to take many more symptoms into account. Indeed apart from the labour concerned there is no reason why all 44 items in the earlier study (Trouton and Maxwell, 1956) might not be included in a prediction equation.

At this stage the reader may have noted an apparent contradiction. When the multiple correlation coefficient was squared and multiplied by 100 we obtained 11·9, and stated that we could, using all three symptoms, predict 11·9 per cent. of the variance of the criterion. Had we used only one symptom, say "d", which correlates to the extent of 0·318 with the criterion and had we squared this correlation and multiplied the answer by 100 we would have obtained 10·1, so that using just this one symptom we would claim to predict a smaller amount of the variance of the criterion than when we used three symptoms. It might appear then that a patient with just this symptom has a poorer prognosis than a patient with all three symptoms, but unless the symptoms interact (Meehl, 1950) to the patient's advantage such a conclusion would be absurd. The dilemma is cleared up if we note that regression weights are not independent of the units used when making our original observations and the regression weights given above refer to a scoring system where the variances of the independent variables are equal and the mean score of each is zero. For example if the presence of a symptom is denoted by -1 then the absence of that symptom should be denoted by $+1$, so that the mean is zero. If now we write out our regression equation as:

$$_{c}=0\cdot206d+0\cdot089a+0\cdot128g$$

where $_{c}$ stands for our estimate of c, and we compare the c-scores of two patients one of whom has all three symptoms while the other has only the

https://doi.org/10.1192/bjp.105.439.463 Published online by Cambridge University Press

second, then the first man's c-score is found to be −0·423, while the second man's score is much higher, namely 0·245. We see now that whereas the amount of the variance of the criterion we can predict depends on the number of symptoms we take into consideration—in general the more the better—a man's actual predicted score on the criterion increases the fewer symptoms he has.

A Factorial Model

Tetrachoric correlations, as we have noted, are calculated from 2×2 contingency tables, but there is an underlying assumption that each pair of variables—though dichotomized—is distributed in a bivariate normal manner. It is known that even slight departure from normality causes the tetrachoric correlations to be extremely unreliable (Kendall, 1949), so that such correlations are not very trustworthy. With a multivariate model, like the multiple regression model, chances of error creeping in by way of tetrachoric correlations are even greater. To avoid the difficulty a more accurate method of dealing with multivariate dichotomous data has been outlined by Dyke and Patterson (1952), and others, and it will be well to look at our data in the light of their recommendations.

The presence (−) or absence (+) of the three symptoms for the 819 patients under consideration can be arranged as in Table I and the proportion of patients in each of the eight resulting categories that recovered can be examined.

Table I

Incidence of Symptoms Amongst Psychiatric Patients

Patients	819							
Depressed (d)	345 (+)				474 (−)			
Anxious (a)	211 (+)		134 (−)		172 (+)		302 (−)	
Delusions (g)	205 (+)	6 (−)	128 (+)	6 (−)	139 (+)	33 (−)	216 (+)	86 (−)
Recovered	68/205	3/6	58/128	3/6	70/139	23/33	129/216	59/86
Proportion recovered	0·3317	0·5000	0·4531	0·5000	0·5036	0·6970	0·5972	0·6860

The data when so arranged call to mind a 2^n factorial design in analysis of variance (Maxwell, 1958), and this suggests how they may be analysed.

A major disadvantage of correlational analyses, as Dyke and Patterson emphasize, is that they do not provide us with estimates of the effects of the separate factors—in our case the separate symptoms—involved in a study. To overcome this difficulty these writers prefer a model in which the effects of the separate factors may safely be taken to be additive and one from which accurate estimates of these effects and of their error variances may be obtained. In brief they want to be able to look at the proportion of patients in each of the eight categories (the last row of Table I) and to say what contribution the presence or absence of a symptom has made to it. Equally important they want to have a check on possible non-linear or interaction effects between the symptoms where recovery is concerned, and the multiple regression model—far from providing us with a test of the existence of such non-linear effects—tacitly assumes that they are either non-existent or unimportant. But let us return to the data in Table I.

Proportions, since by definition they lie between the limits 0 and 1, are not normally distributed. One of the consequences is that they cannot, without suitable transformation, be thought of as being composed of additive components. To put them on an additive scale Dyke and Patterson recommend a

https://doi.org/10.1192/bjp.105.439.463 Published online by Cambridge University Press

logit transformation, though a probit transformation would also have a lot to recommend it. Once the proportions have been transformed and put on an additive scale we can assume, as is the custom in analysis of variance designs, that each is composed of additive effects. In our example it will be convenient in the first place to assume that over and above the general mean of the eight transformed proportions there are three other effects—apart from random error—one due to each of the variables (or symptoms) being considered. It may well be that in addition to these there are other important effects due to interaction which should be considered but the method outlined by Dyke and Patterson furnishes a test of the possibility of the existence of such effects after the main ones have been taken into account.

An important additional point to notice is that whereas the readings in the cells of an ordinary factorial design are each based on the same number of observations, the numbers on which the eight proportions in our investigation are based differ widely. This fact greatly complicates the analysis of the data so that instead of a straightforward analysis of variance providing us with unbiased estimates of the several effects in which we are interested maximum likelihood estimates of these effects have to be obtained by an iterative process. The details of the method are outlined in detail by Dyke and Patterson so that only the results of the analysis of our data need be given here. If we denote by the symbol m′ the maximum likelihood estimate of the overall mean of the eight transformed proportions, and the maximum likelihood estimates of the effects of the three symptoms by the symbols d′, a′ and g′ then these estimates with their standard errors are found to be:

	Main Effect	Statistical Error
m′	·07827	·05338
d′	·19814	·03855
a′	·01030	·03708
g′	·13572	·05288

The estimate of the effect of d′ (depression) is highest, with that for g′ (delusions of guilt) second. These results agree with those found when the partial regression coefficients were obtained. The standard error of g′ is considerably larger than that for either d′ or a′. This was to be expected, for on examination of the frequencies in the fourth row of Table II, it is seen that for non-depressed patients the presence of the symptom "delusions of guilt" was very infrequent, only 12 cases out of 345 showing the symptom, so that an estimate of the effect of the symptom could hardly be expected to be very accurate.

Using the standard errors the estimates of each of the main effects can be compared with each other and with the estimate of the overall mean. The only significant difference obtained when these comparisons are performed is that between d′ and a′ ($P < \cdot 001$). This result bears out our earlier conclusion that prognosis for a patient suffering from depression is somewhat better than that for an anxious patient. The finding that none of the main effects differs significantly from the general mean is in keeping with the rather low correlations between the symptoms and recovery.

The final and perhaps most important stage in the present analysis is the test of the assumption that an additive model, which while allowing for main effects neglects interaction effects, is adequate for accounting for the variation in the data. The appropriate test is a chi-square test: for our data chi-square is found to be 6·639 which, with 4 degrees of freedom, is not significant. We

https://doi.org/10.1192/bjp.105.439.463 Published online by Cambridge University Press

may, therefore, rest assured that no interaction effects of note exist between the symptoms where recovery is concerned.

Conclusions

Two models, a multiple regression model and a factorial model, for use in prediction studies are illustrated by means of a practical example. The first model yields regression weights which indicate the relative importance of the independent variables for predicting the criterion, while the multiple correlation between the combined independent variables and the criterion furnishes a measure of the effectiveness of the battery as a whole for predicting the criterion. The model, however, has disadvantages in that it assumes a multivariate normal distribution for the variables concerned; it provides no check on possible non-linear or interaction effects between the variables nor does it furnish us with independent estimates of the separate effects of the variables. Moreover, in the present example, where the variables are dichotomous so that inefficient statistics such as tetrachoric correlations have to be used there is serious danger of error.

The factorial model, on the other hand, though laborious to use, allows—after a transformation of the data—independent estimates of the effects of the separate variables to be obtained together with estimates of their standard errors, and so makes tests of significance possible. Moreover, it provides us with a check on the existence of interaction effects between the variables when predicting the criterion.

For the sample of patients and the three variables, *depression*, *anxious*, and *delusions of guilt*, considered here it is found that there are no interaction effects where prediction of *recovery* is concerned. As a result we may expect the results given by each model to agree, though prior to a test for interaction we could not have asserted this with confidence. Of the three symptoms prognosis is slightly better for patients who suffer from *depression* alone than for those who suffer from *delusions of guilt*, or *anxiety*, in that order. Prediction on the basis of just these three symptoms, however, is not very reliable for the multiple regression model shows that together they can predict only 12 per cent. of the variance of the criterion.

To conclude it would appear that when multivariate dichotomous variables are concerned a multiple regression model is satisfactory if we are confident that the variables do not interact and if an examination of the 2×2 contingency tables convinces us the distributions are bivariate normal—a quick check on the latter condition is that entries in the diagonally opposite cells of the tables should be approximately equal. When these conditions are met the multiple regression model has the advantage that, though it involves laborious calculations when the number of independent variables is large, these calculations are less laborious than those involved in the *factorial* model. When, however, the assumption of normality is not adequately met and if interaction effects are suspected then the factorial model should be used.

References

Dyke, G. V., and Patterson, H. D., "Analysis of Factorial Arrangements when the Data are Proportions", *Biometrics*, 1952, **8**, 1–12.

Kendall, M. C., "Rank and Product Moment Correlations", *Biometrika*, 1949, **36**, 177–193.

Maxwell, A. E., *Experimental Design in Psychology and the Medical Sciences*, 1958. London: Methuen & Co. Ltd.

Meehl, P. E., "Configural Scoring", *J. Consult. Psychol.*, 1950, **14**, 165–171.

Trouton, D. S., and Maxwell, A. E., "The Relation between Neurosis and Psychosis", *J. Ment. Sci.*, 1956, **102**, 1–21.

A correction to a previous article, by D. S. Trouton and A. E. Maxwell, is given overleaf.

https://doi.org/10.1192/bjp.105.439.463 Published online by Cambridge University Press

A Correction

The signs were omitted from Table V on page 18 of the article: "The Relation between Neurosis and Psychosis" (*J. Ment. Sci.*, 1956, **102**, 1–21). The corrected table is given below, by kind permission of the Editor-in-chief.

Table V

Orthogonal Rotated Factor Loadings

Variables	I	II	III	IV	V	VI	h^2
1	−·142	−·120	−·368	−·123	·118	·499	·448
2	·664	−·493	·176	−·003	−·215	·219	·809
3	·276	−·018	−·058	·056	−·134	·278	·178
4	−·029	·322	·022	·291	·051	−·124	·208
5	−·032	·349	·166	−·119	−·198	·120	·218
6	−·005	·468	·120	·001	−·203	·045	·276
7	·133	·382	·271	·159	−·178	−·302	·385
8	−·019	·308	·433	−·068	−·050	−·034	·291
9	·203	·276	·120	·461	−·114	−·178	·388
10	−·229	·247	−·105	−·237	−·035	·358	·310
11	·014	·483	·107	·097	−·220	−·176	·334
12	−·218	·539	·141	−·222	·095	−·061	·420
13	·120	·373	−·116	·149	−·089	·367	·332
14	−·108	·718	−·210	−·026	−·154	·017	·596
15	·084	·434	−·056	−·220	−·442	·133	·460
16	·019	·296	·389	·187	·068	−·029	·280
17	−·222	·425	·377	·007	−·167	·104	·410
18	·132	·157	·074	·018	−·385	·125	·212
19	·321	·093	·164	·095	−·040	·339	·264
20	·156	−·270	·008	·096	·131	·253	·188
21	−·130	·330	−·147	·037	·500	−·005	·399
22	·594	−·133	−·111	·435	·026	·143	·593
23	·397	−·126	·060	−·595	·101	·021	·542
24	·334	·180	−·350	−·308	·125	·132	·394
25	·556	·024	·202	−·516	−·130	−·081	·640
26	·615	−·002	·086	·533	·050	−·186	·707
27	·315	·378	−·009	·263	·040	·405	·487
28	·612	−·148	·486	−·087	·139	−·033	·661
29	·302	·207	−·616	·355	−·113	·135	·671
30	−·016	·425	−·529	·031	·072	−·094	·476
31	·524	−·008	·372	−·381	−·006	−·169	·587
32	·587	−·004	·034	−·253	−·047	−·246	·472
33	·298	·224	−·031	−·351	−·201	−·214	·349
34	·396	·106	−·334	−·015	−·182	·196	·351
35	·625	−·463	·497	·031	−·014	−·201	·894
36	·539	−·040	·127	·020	·213	−·011	·354
37	·568	·104	−·490	·206	−·030	·057	·620
38	·573	−·275	·135	·088	·033	−·432	·618
39	·667	−·221	·125	−·059	·226	−·191	·600
40	·604	−·242	·254	−·135	·206	−·308	·643
41	·308	·192	−·431	−·176	·176	·339	·494
42	·389	−·010	−·086	·184	·208	·350	·358
43	−·040	·397	−·202	−·010	·532	·021	·484
44	·207	−·230	·579	−·284	−·259	·053	·582
45	−·032	−·163	−·580	·174	−·059	·082	·404

A. E. Maxwell

https://doi.org/10.1192/bjp.105.439.463 Published online by Cambridge University Press