# ON THE USE OF LINEAR DISCRIMINANT FUNCTIONS IN THE REALM OF INDUSTRIAL ACCIDENT INSURANCE

BY

J. VAN KLINKEN
Amsterdam

## 1. INTRODUCTION

Discriminant analysis is an application of multivariate analysis, which may have its use in determining accident risk levels and premiums of industrial enterprises. This paper only aims to give some suggestions. The following questions will be considered.

1. Determining a discriminant function which makes it possible to discriminate between the risk levels of the industrial branches in an efficient way. The industrial branches comprise enterprises with comparable risk levels, hence they are to be considered as homogeneous groups. The function will at the same time serve as a means to classify seperate enterprises into one of these groups.
2. Fixing risk functions which enable us to rank the enterprises of an industrial branch to increasing risk on the ground of observations of a number of variates which characterize the risk situation.
3. Using these risk functions to calculate premiums.

The classification-question mentioned under 1 was the reason to consider the technique of the discriminant analysis. By virtue of the Dutch Industrial Accidents Act every five years a tariff-decree is being published. This decree contains the premiums per wage-unit for the industrial branches. However, there are enterprises e.g. large compound enterprises which do not fall under these regulations. These enterprises ought to be classified according to their own experience. That means we need the knowledge of the risk levels of these particular enterprises in relation to the fixed risk levels of the industrial branches. As mentioned, this is a problem inherent to the typical Dutch situation. It seems, however, probable that such problems and the techniques we intend to

sketch have a wider and more general meaning for the accident insurance.

Discriminant analysis may be looked upon as of a purely algebraical character, just as the least squares method. Stochastic elements in the theory can be completely left out of consideration. This is, in a certain way, advantageous because the well-known procedure of fitting theoretical distributions to the frequency distribution of accident costs may lead to rather bad results. A favourite stochastic model is the compound distribution arising from a simple Poisson-distribution for the number of claims and a $\Gamma$— distribution for the size of the claim. However, the author found that fitting the $\Gamma$— distribution to the observations of Dutch enterprises failed in some cases because, as a small investigation showed, small accidents were frequently not reported by the employers. Rather a certain type of truncated $\Gamma$— distribution might fit the observations, but such a model would be too complicated.

2. *Discriminant analysis*

First we shall try to explain in some detail the algebra involved. The risk situation of an enterprise is characterized by certain variables $x_1, \ldots, x_p$. We consider linear functions $z = \sum_{h=1}^{p} \lambda_h x_h$, where $\lambda_1, \ldots, \lambda_p$ are parameters we have still at our disposal. Let the observations be $x_{hij}$, $h = 1, \ldots, p$ the type of the variable, $i = 1, \ldots, n$ the industrial group and $j = 1, \ldots, n_i$ the number of the enterprise in the group. We write $\bar{x}_{hi} = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{hij}$, $\bar{x}_h = \frac{1}{n} \sum_{i=1}^{n} \bar{x}_{hi}$, $\bar{z}_i = \sum_{h=1}^{p} \lambda_h \bar{x}_{hi}$ etc.

Now the technique consists in determining $\lambda_1 \ldots, \lambda_p$ in such a way that $z$ has very little overlap for the groups. As regards the algebra two cases can be distinguished.

a. We have only to discriminate between two industrial branches (homogeneous groups).

b. The analysis concerns more than two groups.

*Case a.* Clearly the solution is obtained by maximizing the function

$$G = \frac{(\bar{z}_1 - \bar{z}_2)^2}{\sum_{i=1}^{2} \sum_{j=1}^{n_i} (z_{ij} - \bar{z}_i)^2}$$

If we write $\bar{x}_{h1} - \bar{x}_{h2} = d_h$ and

$$\sum_{i=1,2} \sum_{j=1}^{n_i} (x_{hij} - \bar{x}_{hi})\,(x_{h'ij} - \bar{x}_{h'i}) = s_{hh'}$$

we find after some rearrangement

$$(\bar{z}_1 - \bar{z}_2)^2 = \sum_{h,h'} \lambda_h \lambda_{h'}\, d_h d_{h'}$$

$$\sum_{i=1,2} \sum_{j=1}^{n_i} (z_{ij} - \bar{z}_i)^2 = \sum_{h,h'} \lambda_h \lambda_{h'}\, s_{hh'}$$

The system $\dfrac{\partial G}{\partial h'} = 0$, $h' = 1, .., p$ appears to be equivalent with

$$\sum_{h=1}^{p} \lambda_h\, s_{hh'} = c d_{h'}, \qquad h' = 1, ., p$$

where

$$c = \frac{1}{G} \sum_{h=1}^{p} \lambda_h\, d_h$$

Solving this last system we obtain the $\lambda_h$ except for the constant $c$. But as regards the maximum of $G$, only the ratio $\lambda_1 : \lambda_2 : \ldots : \lambda_p$ matters; we may put $\sum_{h=1}^{p} \lambda_h = 1$. With these normalized values we can compute the value $z$ of the enterprise we want to investigate and classify.

Next *case b*. An obvious extension is now to maximize the function

$$G = \frac{\sum_{i=1}^{n} (\bar{z}_i - \bar{z})^2}{\sum_{i=1}^{n} \sum_{j=1}^{n_i} (z_{ij} - \bar{z}_i)^2}$$

Again if we write

$$\sum_{i} (x_{hi} - \bar{x}_h)\,(\bar{x}_{h'i} - \bar{x}_{h'}) = u_{hh'}$$

we get

$$G = \frac{\sum_{h,h'} \lambda_h \lambda_{h'}\, u_{hh'}}{\sum_{h,h'} \lambda_h \lambda_{h'}\, s_{hh'}}$$

The system $\frac{\partial G}{\partial \lambda_{h'}} = 0, h' = 1,., p$ reduces

to $\sum_{h=1}^{p} \lambda_h (u_{hh'} - G s_{hh'}) = 0, h' = 1,., p$

This last system has only a solution if the determinant $| u_{hh'} - G s_{hh'} |$ equals zero. Let us consider the determinantal equation $| u_{hh'} - \mu s_{hh'} | = 0$. Because $u_{hh'} = u_{h'h}, s_{hh'} = s_{h'h}$ the latent roots are real. In practice we obtain a set of different real roots. Hence there is a largest root $\mu_1$. With this $\mu_1$ corresponds the latent vector $(\lambda^*_1,., \lambda^*_p)$. If we replace in the expression for $G$, $G$ by $\mu_1$ and $\lambda_h$ by $\lambda^*_h$, $h = 1,., p$, we see at once that $\lambda^*_h$, $h = 1,., p$ is the solution. We normalize again $\lambda^*_h$ by demanding $\sum_{h=1}^{p} \lambda^*_h = 1$. This method for more than two groups is essentially the method of principal components (2, p. 103). To determine $\mu_1$ and $\lambda^*_h$, $h = 1,., p$, we can make use of special iteration procedures; a practical guide to them can be found a.o. in (2. p.p. 331-358).

So far the sketch of the mathematics connected with this technique. It is clear that in general the function $z$ enables us to discriminate between the groups in an efficient way, because the observed values of $z$ form condensed clouds for the different industrial branches. But there remain open a number of important practical questions. Which variables should be taken into account? There are variables as the observed number of claims, the size of the claims per wage-unit, which we may call properly claim variables. By means of them we can compute exactly the total claim or costs if we know in addition the wage sum. Other variables e.g. the salary level, the average age of the workers, the proportion of man-years spent to specified dangerous work etc., give only indications as regards the risk situation. These variables we may call "risk" variables. Should variables of both types be inserted into the function $z$? From a purely algebraical point of view there are no objections. But already stating that there are "dependent" and „independent" variates gives reason to certain doubts whether mingling these variables into one linear function should be logical and sensible. At first it seems appropriate to base the function exclusively on "claims" or "risk" variables. The choice of the risk variables is rather a difficult one, because we must be sure that the

risk variables are really defining the risk situation. This will become more evident in the following sections. Here we propose to consider only claim variables. Further we demand that the set of variables should completely determine the total observed accident costs. Then we may hope that there is a strong relation between $z$ and the total accident costs. Otherwise the classification procedure may lead to erroneous results. To clarify our intention we take as example the following set of variables.

$N_1$, $a_1$ the number of claims not involving present values, respectively the average size of these claims,
$N_2$, $a_2$ the number of disability pensions and the average size of the present values,
$N_3$, $a_3$ the number of widow pensions and the size of the corresponding present values,
$w$ the number of man-years,
$l$ the amount of insured wages.

We introduce the variables $x_1 = \frac{N_1}{w}$, $x_2 = \frac{N_2}{w}$, $x_3 = \frac{N_3}{w}$, $x_4 = \frac{a_1 w}{l}$, $x_5 = \frac{a_2 w}{l}$, $x_6 = \frac{a_3 w}{l}$ and the discriminant function $z = \sum_{h=1}^{6} \lambda_h x_h$ The total accident costs are

$$S = l\,[x_1 x_4 + x_2 x_5 + x_3 x_6]$$

If we know the values of $x_1, .., x_6$ and $l$, $S$ is fixed. We get a simpler function if we use only $x_1, x_2$ and $x_3$. We may suppose that $N_1, N_2$ and $N_3$ are approximately independent Poisson-variables. It is easily verified that in *case a* the parameters $\lambda_h$ are approximately given by $\lambda^h = \frac{\bar{x}_{h1} - \bar{x}_{h2}}{\bar{x}_{h1} + \bar{x}_{h2}}$, $h = 1, 2, 3$. This is a very simple solution, but a serious drawback is that $x_1$, $x_2$ and $x_3$ do not determine the accident costs completely, the relation with reality is lost. We may classify an enterprise into the group with relative low premium level, whereas its accident costs are very high owing to high values of $a_1$, $a_2$ and $a_3$. Such a classification technique would be senseless.

3. *Canonical variates.*

In the foregoing section we already made the distinction between "claim" and "risk" variables. It is natural to search for a relationship between the two sets of variables. For instance such a relationship may enable us to order some enterprises of an industrial branch only on the ground of the observations of the risk variables.

We consider again linear functions

$$u = \sum_{h=1}^{p'} \lambda_h x_h \text{ and } v = \sum_{h=p'+1}^{p} \lambda_h x_h, \qquad x_1, .., x_p$$

risk variables, $x_{p'+1}, .., x_p$ claim variables. Next we try to construct a risk curve $v(u)$ on the basis of the observations $x_{hj}$, $j = 1, .., n$. It is clear that such a function can be plotted more easily if $v$ and $u$ are strongly correlated. This can be achieved by maximizing

$$G = \sum_{j=1}^{n} (u_j - \bar{u})(v_j - \bar{v})$$

under the conditions

$$\sum_{j=1}^{n} (u_j - \bar{u})^2 = \sum_{j=1}^{n} (v_j - \bar{v}_j)^2 = 1.$$

If we write again

$$\sum_{j=1}^{n} (x_{hj} - \bar{x}_h)(x_{lj} - \bar{x}_l) = s_{h,l} \text{ etc.}$$

we have

$$G = \sum_{h=1,l=p'+1}^{p',p} \lambda_h \lambda_l s_{h,l}$$

($h$ being the risk index, $l$ the claim index).

Application of the method of undetermined multipliers gives

$$\sum_{l=p'+1}^{p} \lambda_l s_{h'l} + \mu \sum_{h=1}^{p'} \lambda_h s_{h',h} = 0, \qquad h' = 1, .., p'$$

$$\sum_{h=1}^{p'} \lambda_h s_{hl'} + \nu \sum_{l=p'+1}^{p} \lambda_l s_{l,l'} = 0, \qquad l' = p' + 1, .., p$$

Multiplying the equations by $\lambda_h, \lambda_l$ and summing the two subsets of equations, we get at once for the values of the multipliers $\mu = \nu = -G$. Treating the factor $G$ as parameter, there results again a homogeneous linear system in the $\lambda_h$. The determinant is a polynomial in $G$ of degree $p$. The vector $(\lambda_1^*, .., \lambda_p^*)$ which corresponds to the largest root $G$ is the solution. We may plot the $(u_j, v_j)$ points and draw a curve. The $v(u)$ graph now enables us to find the ,,$v$'' position of enterprises for which only $x_1, .., x_p$ are known.

The technique briefly sketched is the method of canonical correlation, $u$ and $v$ are called canonical variates (2, pp. 114-121). As in the case of discriminant functions there are difficulties in explaining the curve $v(u)$. In the first place we have only $n$ values $v(u)$, but are we indeed justified to speak of a function, especially a one valued function $v(u)$? Another question concerns the correlation of $v$ and the total costs. In case of a weak correlation $v$ lacks any actuarial meaning. We here see that both the methods of discriminant analysis and canonical correlation can be used for classification purposes of accident risks but that it remains somewhat doubtful if the procedures can always be sensibly interpreted by the actuary. This is of course a serious drawback. Therefore in section 4 we shall propose a few modifications which imply perhaps a certain improvement.

## 4. *Calculation of premiums.*

We need a coherent system of classification and determination of premiums. The discriminant function $z$ and the canonical variate $v$ derived as indicated in the sections 2 and 3 are not quite suited to this end. In the foregoing sections, especially at the end of section 3, we noted this already. Without going further into details which imperfections adhere to these methods, we shall now briefly describe how to try to overcome the difficulties in some degree by modifying and combining both techniques.

We propose to introduce linear cost functions.

$p = \sum_{l=1}^{q} \sigma_l x_l$, where the $x_l$ are costs variables.

By cost variables we mean the sum of net costs of claims of a specified type per wage unit. The terms $\sigma_l x_l$ may represent gross-costs

per wage unit. Further, we shall distinguish between risk variables which characterize the industrial group as such and risk variables which mark the individuality of the enterprises within the groups. The group risk variables may be $x_h$, $h = 1,.., p'$ the enterprise risk variables $x_h$, $h = p' + 1,.., p$. On the ground of observations of a recent period we determine the maxime of

$$G_1 = \sum_{i=1}^{n} (\bar{p}_i - \bar{p})(\bar{u}_i - \bar{u})$$

under the condition $\frac{1}{n}\sum_{i=1}^{n} (\bar{u}_i - \bar{u})^2 = 1$

($i$ indicates the group, $\bar{p}_i = \sum_{l=1}^{q} \sigma_l \bar{x}_{li}, \bar{u} = \sum_{h=1}^{p'} \lambda_h \bar{x}_{hi}$ etc.). Following again Lagrange's method of multipliers there result the maximum conditions

$$\sum_{l=1}^{q} \sigma_l s_{h',l} + \mu \sum_{h=1}^{p'} \lambda_h s_{h',h} = 0, \qquad h' = 1,., p'$$

$$\sum_{h,h'=1}^{p'} \lambda_h \lambda_{h'} s_{hh'} = 1$$

$$s_{h,l} = \frac{1}{n}\sum_{i=1}^{n} (\bar{x}_{hi} - \bar{x}_h)(\bar{x}_{li} - \bar{x}_l) \text{ etc.}$$

Again it is easily verified that $\mu = -G$. Treating $\frac{1}{G}$ as a constant $c$ we get the linear system of inhomogeneous equations

$$\sum_{h=1}^{p'} \lambda_h s_{h'h} = c \sum_{l=1}^{q} \sigma_l s_{h'l}, \qquad h' = 1,., p'$$

The solution is determined except for an unknown factor. In addition this factor can be found from the subsidiary condition

$$\sum_{h,h'=1}^{p'} \lambda_h \lambda_{h'} s_{hh'} = 1$$

Hence we obtain a function $u = \sum_{h=1}^{p'} \lambda_h^* x_h$.

Now we form the linear discriminant function

$$z = \sum_{h=1}^{p'} \lambda_h^* x_h + \sum_{h=p'+1}^{p} \lambda_h x_h$$

Only the second term contains parameters $\lambda_h$ we have still at our disposal. Maximizing $G_2 = \dfrac{\sum_{i=1}^{n} (\bar{z}_i - \bar{z})^2}{\sum_{i,j=1}^{n,n_i} (z_{ij} - \bar{z}_i)^2}$ leads to the equations

$$\sum_{h=1}^{p'} \lambda_h^* (u_{hh'} - G_2 s_{hh'}) + \sum_{h=p'+1}^{p} \lambda_h (u_{hh'} - G_2 s_{hh'}) = 0$$

$$h' = p' + 1, .., p$$

(see section 2)

Anew treating $G_2$ as un unknown factor we have again a system of inhomogeneous linear equations. We may solve them formally and obtain the system

$$\lambda_{h'} = f_{h'} (\lambda_{p'} + 1, .., \lambda_p), \; h' = p' + 1, .., p$$

If in the $\lambda$-region considered

$$\sum_{h'=p'+1}^{p} \left| \frac{\partial f_{h'}}{\partial \lambda_h} \right| \leq M < 1 \quad h = p' + 1, .., p$$

we can apply the well-known iteration procedure (3). Let us assume that by iteration we have found the solution $(\lambda^*_{p'+1}, .., \lambda^*_p)$. The discriminant function $z = \sum_{h=1}^{p} \lambda_h^* x_h$ is now completely fixed. Next we determine functions $v_i = \sum_{h=p'+1}^{p} \mu_{hi} x_h$ by maximizing $G_{2,i} = \sum_{j=1}^{n_i} (p_{ij} - \bar{p}_i)(v_{ij} - \bar{v}_i)$ under the condition $\frac{1}{n_i} \sum_{j=1}^{n_i} (v_{ij} - \bar{v}_i)^2 = 1$.

Following the algebra already given this leads to functions

$$v_i = \sum_{h=p'+1}^{p} \mu_{hi}^* x_h$$

We graduate the observed points $(p_{ij}, v_{ij})$ and obtain graph's $p_i(v)$ we may consider as premium curves. Once the functions $z$ and $p_i(v)$ are obtained, we can handle a "new" enterprise for which only the

values of the risk variables are available. First compute its value $z$ and classify the enterprise into one of the groups. Let this group have number $i$. Next compute its $v$-value. From the $p_i(v)$ graph we finally may derive the premium.

Finally we wish to draw attention to the fact that we may use $z$ and $v$ as statistics in tests if we estimate the variances and the covariances of the $x_h$, $h = 1,., p$. If $H_o$ means the hypothesis that the enterprise belongs to a certain group $i$ we might verify this by comparing

$$\tilde{z}_i = \frac{z - \bar{z}}{\sqrt{\sum_{h=1}^{p} \lambda_h^{*2} \operatorname{Var} x_{hi} + 2 \sum_{h,h'=1}^{p} \lambda_h^* \lambda_{h'}^* \operatorname{Cov}(x_{hi}, x_{h'i})}}.$$

with the table of the $N(0, 1)$ — distribution. Of course the procedure here proposed is laborious, but, maybe, it can help the actuary confronted with certain questions as regards classifying risks involving several factors. As the foregoing gives only a suggestion, actual numerical evolutions are needid in order to be able rating the method at its true value.

## LITERATURE

1. P. HOEL, Introduction to mathematical statistics, New York 1946, pp. 121-126.
2. G. TINTNER, Econometrics, John Wiley & Sons, Inc., New York, chapter 6.
3. F. A. WILLERS, Practical Analysis, Dover publications 1947, pp. 212, 213.
4. V. N. FADDEEVA, Computational methods of linear algebra, Dover publications, 1959.