Discriminant analysis (a type of regression)

1. Step: Preprocessing: Transform the nonlinear world of benefits for illness to a linear framework

2. Step: Discriminant analysis

- Build clusters of values (groups) of the random varable Y (cost of "linearized" illnes), that is to be predicted
- The random variables X_i (representing the verious diseases, i=1,...,15,000) will predict Y
- Modelling the total (linearized) cost of illnes using "discriminant parameters b_i":

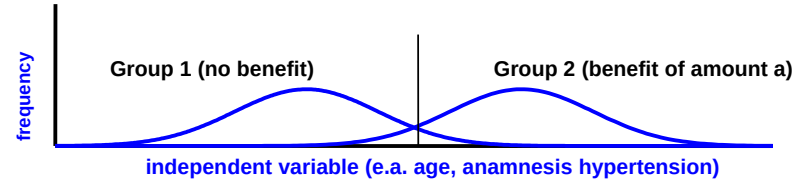
$$Y_D = b_0 + b_1X_1 + b_2X_2 + + b_mX_m \qquad \text{(Discriminant function)}$$

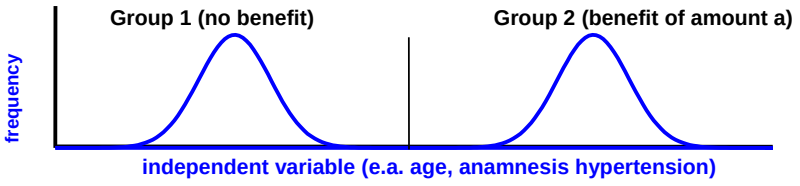
- determine b_i so, such that:
  the distance of the mean values of the groups is maximal
  the variance within the groups is minimal
- Select adequate predictors X_i such that the loss of information is minimal.

3. Step: Postprocessing Return to the nonlinear world of cost of illnes by adjusting functions

Goals of the selection of coefficients

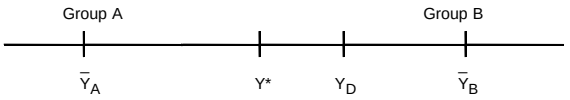
- **large distance between the "centroids" (mean value of the considered group)**
- **small variance witin groups**
- **large variance between groups**
- **metric is the Mahalanobis distance**





DECISIONS by DISCRIMINANT FUNCTION

- $Y_D = b_0 + b_1X_1 + b_2X_2 + \ldots + b_mX_m$ (Discriminant function)
- evaluate dicriminant function for known customers with special values of the predictors X_i
- the discriminant functions' values are assigned to points on the discriminant axis.
- compare the values with the group -centroids.



$\bar{Y}_A$, $\bar{Y}_B$ =Group-centroids = mean value of discriminant function Y_D for thegroup A or B respectively.
Y* = criterion of separation

- evaluate Y_D for new customers with special values of the predictors X_i according to the value of separation Y* of the group A or B

SELECTION of SIGNIFICANT PREDICTORS for DISCRIMINANT ANALYSIS

- selection of significant predictors is realized step by step .
  select an additional variable in every step, such that discrimination is optimized between groups with the smallest (Mahalanobis-) distance (F to enter)
- if variables contribute only insignificantly to discrimination in the current multivariate system, they can be removed in a later step
- measure the quality of dicriminance by Wilks Lambda λ: $\lambda = \frac{\text{non explained scattering}}{\text{total scattering}}$
- the level of significance of the discriminant analysis results from λ as a χ^2-distributed variable.

$$\chi^2 = -\left(N - \frac{J + G}{2} - 1\right) \ln(\lambda)$$

N = number of customers in the sample, J = number of predictors G = number of groups

- the level of significance is better (lower) than α= 0.001 for the present study

DISCRIMINANCE BETWEEN 2 GROUPS

group 1 = no benefit payed group 2 = benefit payed i. e. G = 2 (2 groups)

$X_{kl}^{(i)}$ = random variable, describing the predictor number k of the person number l,

who belongs to the with benefits (i), $l=1...n_i$; $n = n_1 + n_2$

$Y_l^{(i)}$ = random variable, describing the "probability to benefits" of person l belonging to group i

$Y_l^{(1)} := b_1X_{1l}^{(1)} + b_2X_{2l}^{(1)} + \dots + b_mX_{ml}^{(1)}$ group (1)

$Y_l^{(2)} := b_1X_{1l}^{(2)} + b_2X_{2l}^{(2)} + \dots + b_mX_{ml}^{(2)}$ group (2)

group centroids:

$$\bar{Y}^{(1)} := \frac{1}{n_1}\sum_{l=1}^{n_1} Y_l^{(1)} \qquad \bar{Y}^{(2)} := \frac{1}{n_2}\sum_{l=1}^{n_2} Y_l^{(2)}$$

$\bar{X}_k^{(i)}$ = mean value of the predictor k in group i

CRITERION of DISCRIMINANCE

$S = S(b_1, \dots, b_m) = \bar{Y}^{(1)} - \bar{Y}^{(2)}$ = distance of the centre of the group = **max!**

$$T = T(b_1, \dots, b_m) = \sum_{l=1}^{n_1} (Y_l^{(1)} - \bar{Y}^{(1)})^2 + \sum_{l=1}^{n_2} (Y_l^{(2)} - \bar{Y}^{(2)})^2$$ = variance between groups = **min!**

$$Q = \frac{S^2}{T} = \textbf{max!} \quad \Rightarrow \quad \frac{\partial Q}{\partial b_k} = 0$$ under minor condition . $\frac{S}{T} = 1$

Solution $\hat{b}_1, \dots, \hat{b}_m$ is inserted in S

$$S(\hat{b}_1, \dots, \hat{b}_m) = (n - g) \sum_{i,j=1}^{m} w_{ij}^* (\bar{X}_i^{(1)} - \bar{X}_i^{(2)})(\bar{X}_j^{(1)} - \bar{X}_j^{(2)})$$

w_{ij}^* = elements of inversion of the co variance-matrix (within-groups)

$D^2 = S$ is called Mahalanobis - distance

Criterion of selection of significant predictors

- not all predictors X_i are considered immediately.
  search for predictors, that separate the groups optimally

- for that purpose: calculate S depending on only some of the b_i

- first select such X_i (or b_i respectively) , maximizing the distance S between groups, that are nearest

- $F = \frac{(n - 1 - m)\, n_1 n_2}{m\,(n - 2)\,(n_1 + n_2)}\, S$ is F-distributed.