Fisher’s approach to classification with two populations

- Fisher’s idea was to transform the multivariate variables \(X_1, \ldots, X_p\) to a univariate variable \(Y\), which is a linear function of the \(X\) variables, i.e.,

\[
Y = a_1X_1 + a_2X_2 + \cdots + a_pX_p,
\]

such that the \(Y\) observations derived from the two populations were separated as much as possible.

- A fixed linear combination of the \(x\)'s takes the values \(y_{11}, y_{12}, \ldots, y_{1n_1}\) for the observations from the first population and the values \(y_{21}, y_{22}, \ldots, y_{2n_2}\) for the observations from the second population. The separation of these two sets of univariate \(y\)'s is assessed in terms of the difference between \(\bar{y}_1\) and \(\bar{y}_2\), expressed in standard deviation units. That is,

\[
\text{separation} = \frac{|\bar{y}_1 - \bar{y}_2|}{s_y}, \quad \text{where } s_y^2 = \frac{n_1 \sum (y_{ij} - \bar{y}_1)^2 + n_2 \sum (y_{2j} - \bar{y}_2)^2}{n_1 + n_2 - 2}
\]

is the pooled estimate of the variance.

**Result 11.3.** The linear combination \(\hat{y} = \hat{a}'x = (\bar{x}_1 - \bar{x}_2)'S_{\text{pooled}}^{-1}x\) maximizes the ratio

\[
\frac{\text{squared distance}}{\text{between sample means of } y} = \frac{(\bar{y}_1 - \bar{y}_2)^2}{s_y^2} = \frac{(\hat{a}'\bar{x}_1 - \hat{a}'\bar{x}_2)^2}{\hat{a}'S_{\text{pooled}}\hat{a}} = \frac{(\hat{a}'d)^2}{\hat{a}'S_{\text{pooled}}\hat{a}}
\]

over all possible coefficient vectors \(\hat{a}\) where \(d = (\bar{x}_1 - \bar{x}_2)\). The maximum of the ratio is \(D^2 = (\bar{x}_1 - \bar{x}_2)'S_{\text{pooled}}^{-1}(\bar{x}_1 - \bar{x}_2)\).

proof.

- allocation rule based on Fisher’s discriminant function

Allocate \(x_0\) to \(\pi_1\) if \(\hat{y}_0 = (\bar{x}_1 - \bar{x}_2)'S_{\text{pooled}}^{-1}x_0 \geq \hat{m} = \frac{1}{2}(\bar{x}_1 - \bar{x}_2)'S_{\text{pooled}}^{-1}(\bar{x}_1 + \bar{x}_2)\)

Allocate \(x_0\) to \(\pi_2\) if \(\hat{y}_0 < \hat{m}\)

- Note. Fisher’s linear discriminant function was developed under the assumption that the two population, whatever their form, have a common covariance matrix.
Is classification a good idea for your data?

- For two populations, the maximum relative separation that can be obtained by considering linear combinations of the multivariate observations is equal to the distance $D^2$.

- Note. $D^2$ can be used to test whether the population means differ significantly (Hotelling’s $T^2$ test).

  ⇒ a test for differences in mean vectors can be viewed as a test for the “significance” of the separation that can be achieved.

- Note. Significant separation does not necessarily imply good classification. By contrast, if the separation is not significant, the search for a useful classification rule will probably prove fruitless.

- Classification of Normal Populations When $\Sigma_1 \neq \Sigma_2$

\[
f(x) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)}
\]

**Result 11.4.** Let the populations $\pi_1$ and $\pi_2$ be described by multivariate normal densities with mean vectors and covariance matrices $\mu_1, \Sigma_1$ and $\mu_2, \Sigma_2$, respectively. The allocation rule that minimizes the expected cost of misclassification is given by

\[
\begin{align*}
R_1: & \quad -\frac{1}{2} x'(\Sigma_1^{-1} - \Sigma_2^{-1}) x + (\mu_1'\Sigma_1^{-1} - \mu_2'\Sigma_2^{-1}) x - k \geq \ln \left[ \frac{c(112)}{c(211)} \frac{p_2}{p_1} \right] \\
R_2: & \quad -\frac{1}{2} x'(\Sigma_1^{-1} - \Sigma_2^{-1}) x + (\mu_1'\Sigma_1^{-1} - \mu_2'\Sigma_2^{-1}) x - k < \ln \left[ \frac{c(112)}{c(211)} \frac{p_2}{p_1} \right]
\end{align*}
\]

where

\[
k = \frac{1}{2} \ln \left( \frac{|\Sigma_1|}{|\Sigma_2|} \right) + \frac{1}{2} \left( \mu_1'\Sigma_1^{-1}\mu_1 - \mu_2'\Sigma_2^{-1}\mu_2 \right)
\]

- it is called quadratic classification because of the quadratic term

\[-\frac{1}{2} x'(\Sigma_1^{-1} - \Sigma_2^{-1}) x\]

- Quadratic classification rule

  Allocate $x_0$ to $\pi_1$ if

\[
-\frac{1}{2} x_0'(S_1^{-1} - S_2^{-1}) x_0 + (\bar{x}_1'S_1^{-1} - \bar{x}_2'S_2^{-1}) x_0 - k \geq \ln \left[ \frac{c(112)}{c(211)} \frac{p_2}{p_1} \right]
\]

  Allocate $x_0$ to $\pi_2$ otherwise.

- Evaluating classification functions

  - one important way of judging the performance of any classification procedure is to calculate its “error rate,” or misclassification probabilities.

  - total probability of misclassification (population)

\[
\text{TPM} = p_1 \int_{R_2} f_1(x) \, dx + p_2 \int_{R_1} f_2(x) \, dx
\]

  - actual error rate (sample)

\[
\text{AER} = p_1 \int_{R_2} f_1(x) \, dx + p_2 \int_{R_1} f_2(x) \, dx
\]

  - apparent error rate (do not dependent on population densities)

\[
\text{APER} = \text{proportion of items in the training set that are misclassified}
\]
it is easy to calculate and can be calculated for any classification procedure

- it tends to underestimate the AER because the data used to build the classification function are also used to evaluate it

- one procedure is to split the total sample into a training sample and a validation sample, but it required large sample and the information in the validation sample is not used to construct the classification function

- cross-validation method (leave-one-out method)

1. Start with the \( \pi_1 \) group of observations. Omit one observation from this group, and develop a classification function based on the remaining \( n_1 - 1, n_2 \) observations.

2. Classify the “holdout” observation, using the function constructed in Step 1.

3. Repeat Steps 1 and 2 until all of the \( \pi_1 \) observations are classified. Let \( n_{1M}^{(H)} \) be the number of holdout (H) observations misclassified in this group.

4. Repeat Steps 1 through 3 for the \( \pi_2 \) observations. Let \( n_{2M}^{(H)} \) be the number of holdout observations misclassified in this group.

\[
\hat{P}(2|1) = \frac{n_{1M}^{(H)}}{n_1} \\
\hat{P}(1|2) = \frac{n_{2M}^{(H)}}{n_2}
\]

\[
\hat{E}(\text{AER}) = \frac{n_{1M}^{(H)} + n_{2M}^{(H)}}{n_1 + n_2}
\]

- Reading: textbook, 11.1, 11.2, 11.3, 11.4