# Discriminant Analysis
# Description of Group Separation

## 1 Introduction

Discriminant analysis is a classical multivariate technique designed to **describe and measure separation between groups**. It addresses questions such as:

- How can we best distinguish between groups based on observed variables?
- What linear combinations of variables provide the clearest separation?

Suppose we have data on multiple variables for individuals who belong to known groups (e.g., patients classified as "diseased" vs. "healthy," or students from different majors). Although the groups themselves are known, the interest is in *understanding* and *describing* the way variables contribute to group differences.

The key tool is the **discriminant function**, which is a linear combination of variables constructed to maximize group separation. Unlike classification analysis, which focuses on *predicting* group membership for new observations, discriminant analysis is primarily **descriptive**: it summarizes group differences in terms of weighted combinations of variables.

## 2 The Discriminant Function for Two Groups

Consider two groups, labeled Group 1 and Group 2, each with multivariate observations on the same set of variables. Suppose we have:

- $n_1$ observations from Group 1 and $n_2$ from Group 2,
- Each observation is a $p$-dimensional vector of variables,
- Group sample means are $\bar{\mathbf{x}}_1$ and $\bar{\mathbf{x}}_2$,
- Within-group covariance matrices are assumed equal: $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \mathbf{\Sigma}$.

The **linear discriminant function** is defined as:

$$y = \mathbf{a}'\mathbf{x},$$

where **a** is a vector of coefficients chosen to maximize the separation between the two groups.

*Criterion for Separation*

The discriminant function seeks to maximize the ratio of between-group variance to within-group variance in the projected space. The discriminant function transforms the observation vectors into scalars. We find the grouo means of these scalars and choose vector a that maximizes the standardized difference of these means. Specifically,

$$\text{Maximize} \quad \frac{(\mathbf{a}'(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2))^2}{\mathbf{a}'\mathbf{S}\mathbf{a}},$$

where **S** is the pooled within-group covariance matrix.

The solution is proportional to

$$\mathbf{a} \propto \mathbf{S}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2).$$

Thus the maximizing vector **a** is not unique. However, its "direction" is unique; that is, the relative values or ratios of $a_1, a_2, ..., a_p$ are unique, and $z = a'x$ projects points $x$ onto the line on which $(\bar{y}_1 - \bar{y}_2)^2/s_z^2$ is maximized. Note that in order for $S^{-1}$ to exist, we must have $n_1 + n_2 - 2 > p$.

Thus, the discriminant function assigns larger weights to variables that (i) contribute strongly to group mean differences and (ii) have smaller within-group variance.

The optimum direction given by **a** is effectively parallel to the line joining $\bar{x}_1$ and $\bar{x}_2$, because the squared distance $(\bar{y}_1 - \bar{y}_2)^2/s_z^2$ is equivalent to the standardized distance between $\bar{x}_1$ and $\bar{x}_2$. Any other direction than that represented by a would yield a smaller difference between $\mathbf{a}'\bar{y}_1 - \mathbf{a}'\bar{y}_2$.

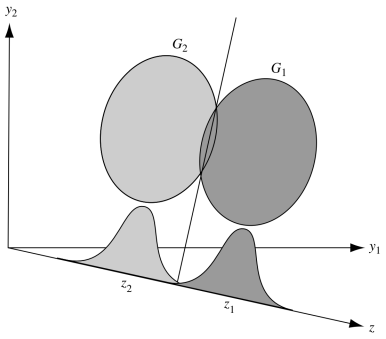


**Figure 8.1.** Two-group discriminant analysis.

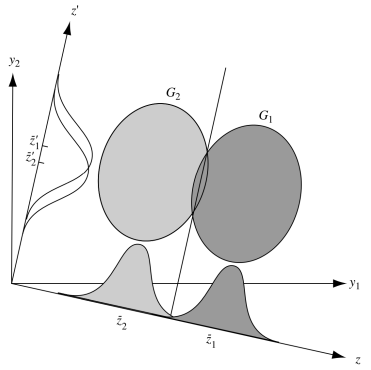


**Figure 8.2.** Separation achieved by the discriminant function.

# 3 Relationship between Two-Group Discriminant Analysis and Multiple Regression

There is a close connection between two-group discriminant analysis and multiple regression. If we code group membership as a binary variable (say, 0 for Group 1 and 1 for Group 2), then regressing this coded variable on the predictors yields coefficients proportional to those from the discriminant function.

Key parallels:

- Regression perspective: predicts group membership using least squares.
- Discriminant analysis perspective: finds linear combinations maximizing group mean separation relative to within-group variability.

Thus, discriminant analysis can be seen as a special case of regression with categorical outcomes, but framed in terms of separation rather than prediction.

# 4 Discriminant Analysis for Several Groups

When more than two groups exist, the approach generalizes. Suppose there are $g$ groups, each with mean vector $\bar{\mathbf{x}}_k$ $(k = 1, 2, \ldots, g)$.

*The General Discriminant Function*

We seek linear combinations of variables,

$$y_j = \mathbf{a}_j' \mathbf{x}, \quad j = 1, 2, \ldots, m,$$

where $m = \min(p, g - 1)$, such that these combinations maximize group separation.

*Between-Group and Within-Group Matrices*

Define:

Within-group scatter matrix:

$$\mathbf{W} = \sum_{k=1}^{g} \sum_{i=1}^{n_k} (\mathbf{x}_{ik} - \bar{\mathbf{x}}_k)(\mathbf{x}_{ik} - \bar{\mathbf{x}}_k)'.$$

Between-group scatter matrix:

$$\mathbf{B} = \sum_{k=1}^{g} n_k (\bar{\mathbf{x}}_k - \bar{\mathbf{x}})(\bar{\mathbf{x}}_k - \bar{\mathbf{x}})',$$

where $\bar{\mathbf{x}}$ is the overall mean.
The goal is to maximize

$$\frac{\mathbf{a}'\mathbf{B}\mathbf{a}}{\mathbf{a}'\mathbf{W}\mathbf{a}}.$$

This leads to solving the generalized eigenvalue problem:

$$\mathbf{B}\mathbf{a} = \lambda \mathbf{W}\mathbf{a}.$$

- Eigenvalues $\lambda$ measure discriminatory power of each function.
- Corresponding eigenvectors $\mathbf{a}$ give discriminant coefficients.

*Number of Discriminant Functions*
At most $g - 1$ discriminant functions can be formed. These represent orthogonal dimensions along which the groups are maximally separated.

# 5 Standardized Discriminant Functions

Because discriminant coefficients depend on variable scaling, it is common to standardize variables. Standardization ensures comparability of coefficients by expressing each variable in units of standard deviation.

The standardized discriminant coefficients are obtained by multiplying raw coefficients by the standard deviations of the variables. They provide clearer interpretation of variable contributions to group separation.

# 6 Tests of Significance

Statistical tests assess whether the observed separation between groups is significant.

*Two-Group Case*
For two groups, the test reduces to Hotelling's $T^2$, which evaluates whether the group mean vectors differ significantly:

$$T^2 = \frac{n_1 n_2}{n_1 + n_2} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' \mathbf{S}^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2).$$

The discriminant function coefficient vector $\mathbf{a}$ is significantly different from 0 if $T^2$ is significant. This statistic is related to an $F$-distribution.

*Several-Group Case*

In Section 8.4.1 we noted that the discriminant criterion $\lambda = \mathbf{a}'\mathbf{H}\mathbf{a}/\mathbf{a}'\mathbf{E}\mathbf{a}$ is maximized by $\lambda_1$, the largest eigenvalue of $\mathbf{E}^{-1}\mathbf{H}$, and that the remaining eigenvalues $\lambda_2, \ldots, \lambda_s$ correspond to other discriminant dimensions. These eigenvalues are the same as those in the Wilks $\Lambda$-test in (6.14) for significant differences among mean vectors,

$$\Lambda_1 = \prod_{i=1}^{s} \frac{1}{1+\lambda_i}, \tag{8.18}$$

which is distributed as $\Lambda_{p,k-1,N-k}$, where $N = \sum_i n_i$ for an unbalanced design or $N = kn$ in the balanced case. Since $\Lambda_1$ is small if one or more $\lambda_i$'s are large, Wilks' $\Lambda$ tests for significance of the eigenvalues and thereby for the discriminant functions. The $s$ eigenvalues represent $s$ dimensions of separation of the mean vectors $\overline{\mathbf{y}}_1, \overline{\mathbf{y}}_2, \ldots, \overline{\mathbf{y}}_k$. We are interested in which, if any, of these dimensions are significant. In the context of discriminant functions, Wilks' $\Lambda$ is more useful than the other three MANOVA test statistics, because it can be used on a subset of eigenvalues, as we see shortly.

In addition to the exact test provided by the critical values for $\Lambda$ found in Table A.9, we can use the $\chi^2$-approximation for $\Lambda_1$ given in (6.16), with $\nu_E = N - k = \sum_i n_i - k$ and $\nu_H = k - 1$:

$$\begin{aligned} V_1 &= -\big[\nu_E - \tfrac{1}{2}(p - \nu_H + 1)\big] \ln \Lambda_1 \\ &= -\big[N - 1 - \tfrac{1}{2}(p + k)\big] \ln \prod_{i=1}^{s} \frac{1}{1 + \lambda_i} \\ &= \big[N - 1 - \tfrac{1}{2}(p + k)\big] \sum_{i=1}^{s} \ln(1 + \lambda_i), \end{aligned} \tag{8.19}$$

which is approximately $\chi^2$ with $p(k - 1)$ degrees of freedom. The test statistic $\Lambda_1$ and its approximation (8.19) test the significance of all of $\lambda_1, \lambda_2, \ldots, \lambda_s$. If the test leads to rejection of $H_0$, we conclude that at least one of the $\lambda$'s is significantly different from zero, and therefore there is at least one dimension of separation of mean vectors. Since $\lambda_1$ is the largest, we are only sure of its significance, along with that of $z_1 = \mathbf{a}_1' \mathbf{y}$.

To test the significance of $\lambda_2, \lambda_3, \ldots, \lambda_s$, we delete $\lambda_1$ from Wilks' $\Lambda$ and the associated $\chi^2$-approximation to obtain

$$\Lambda_2 = \prod_{i=2}^{s} \frac{1}{1 + \lambda_i}, \tag{8.20}$$

$$V_2 = -\left[N - 1 - \tfrac{1}{2}(p+k)\right] \ln \Lambda_2 = \left[N - 1 - \tfrac{1}{2}(p+k)\right] \sum_{i=2}^{s} \ln(1+\lambda_i), \quad (8.21)$$

which is approximately $\chi^2$ with $(p-1)(k-2)$ degrees of freedom. If this test leads to rejection of $H_0$, we conclude that at least $\lambda_2$ is significant along with the associated discriminant function $z_2 = \mathbf{a}_2'\mathbf{y}$. We can continue in this fashion, testing each $\lambda_i$ in turn until a test fails to reject $H_0$. (To compensate for making several tests, an adjustment to the $\alpha$-level of each test could be made as in procedure 2, Section 5.5.) The test statistic at the $m$th step is

$$\Lambda_m = \prod_{i=m}^{s} \frac{1}{1+\lambda_i}, \quad (8.22)$$

which is distributed as $\Lambda_{p-m+1,k-m,N-k-m+1}$. The statistic

$$\begin{aligned} V_m &= -\left[N - 1 - \tfrac{1}{2}(p+k)\right] \ln \Lambda_m \\ &= \left[N - 1 - \tfrac{1}{2}(p+k)\right] \sum_{i=m}^{s} \ln(1+\lambda_i) \end{aligned} \quad (8.23)$$

has an approximate $\chi^2$-distribution with $(p-m+1)(k-m)$ degrees of freedom. In some cases, more $\lambda$'s will be statistically significant than the researcher considers to be of practical importance. If $\lambda_i / \sum_j \lambda_j$ is small, the associated discriminant function may not be of interest, even if it is significant.

We can also use an $F$-approximation for each $\Lambda_i$. For

$$\Lambda_1 = \prod_{i=1}^{s} \frac{1}{1+\lambda_i},$$

we use (6.15), with $v_E = N - k$ and $v_H = k - 1$:

$$F = \frac{1 - \Lambda_1^{1/t}}{\Lambda_1^{1/t}} \frac{\mathrm{df}_2}{\mathrm{df}_1}, \quad (8.24)$$

where

$$t = \sqrt{\frac{p^2(k-1)^2 - 4}{p^2 + (k-1)^2 - 5}}, \qquad w = N - 1 - \tfrac{1}{2}(p+k),$$

$$\mathrm{df}_1 = p(k-1), \qquad \mathrm{df}_2 = wt - \tfrac{1}{2}[p(k-1) - 2].$$

For

$$\Lambda_m = \prod_{i=m}^{s} \frac{1}{1+\lambda_i}, \qquad m = 2, 3, \ldots, s,$$

we use

$$F = \frac{1 - \Lambda_m^{1/t}}{\Lambda_m^{1/t}} \frac{\mathrm{df}_2}{\mathrm{df}_1} \tag{8.25}$$

with $p - m + 1$ and $k - m$ in place of $p$ and $k - 1$:

$$\begin{aligned} t &= \sqrt{\frac{(p-m+1)^2(k-m)^2 - 4}{(p-m+1)^2 + (k-m)^2 - 5}}, \\ w &= N - 1 - \tfrac{1}{2}(p+k), \\ \mathrm{df}_1 &= (p-m+1)(k-m), \\ \mathrm{df}_2 &= wt - \tfrac{1}{2}[(p-m+1)(k-m) - 2]. \end{aligned}$$

# 7 Interpretation of Discriminant Functions

Once discriminant functions are extracted, interpretation focuses on identifying which variables drive group separation. Several tools are available:

1. Standardized coefficients - show relative contribution of each variable, adjusting for scale.

   To offset differing scales among the variables, the discriminant function coefficients can be standardized using (8.16) or (8.17), in which the coefficients have been adjusted so that they apply to standardized variables. For the observations in the first of two groups, for example, we have by (8.15),

   $$\begin{aligned} z_{1i} &= a_1^* \frac{y_{1i1} - \overline{y}_{11}}{s_1} + a_2^* \frac{y_{1i2} - \overline{y}_{12}}{s_2} + \cdots + a_p^* \frac{y_{1ip} - \overline{y}_{1p}}{s_p}, \\ & i = 1, 2, \ldots, n_1. \end{aligned}$$

   The standardized variables $(y_{1ir} - \overline{y}_{1r})/s_r$ are scale free, and the standardized coefficients $a_r^* = s_r a_r$, $r = 1, 2, \ldots, p$, therefore correctly reflect the joint contribution of the variables to the discriminant function $z$ as it maximally separates the groups. For the case of several groups, each discriminant function coefficient vector $\mathbf{a} = (a_1, a_2, \ldots, a_p)'$ is an eigenvector of $\mathbf{E}^{-1}\mathbf{H}$, and as such, it takes into account the sample correlations among the variables as well as the influence of each variable in the presence of the others.

   As noted in Section 8.5, this standardization is carried out for each of the $s$ discriminant functions. Typically, each will have a different interpretation; that is, the pattern of the coefficients $a_r^*$ will vary from one function to another.

2. Partial $F$-values - test whether each variable adds significantly to discrimination beyond others.

   For any variable $y_r$, we can calculate a partial $F$-test showing the significance of $y_r$ after adjusting for the other variables, that is, the separation provided by $y_r$ in addition to that due to the other variables. After computing the partial $F$ for each variable, the variables can then be ranked.

   In the case of two groups, the partial $F$ is given by (5.32) as

$$F = (\nu - p + 1)\frac{T_p^2 - T_{p-1}^2}{\nu + T_{p-1}^2}, \tag{8.26}$$

   where $T_p^2$ is the two-sample Hotelling $T^2$ with all $p$ variables, $T_{p-1}^2$ is the $T^2$-statistic with all variables except $y_r$, and $\nu = n_1 + n_2 - 2$. The $F$-statistic in (8.26) is distributed as $F_{1,\nu-p+1}$.

   For the several-group case, the partial $\Lambda$ for $y_r$ adjusted for the other $p - 1$ variables is given by (6.128) as

$$\Lambda(y_r|y_1, \ldots, y_{r-1}, y_{r+1}, \ldots, y_p) = \frac{\Lambda_p}{\Lambda_{p-1}}, \tag{8.27}$$

   where $\Lambda_p$ is Wilks' $\Lambda$ for all $p$ variables and $\Lambda_{p-1}$ involves all variables except $y_r$. The corresponding partial $F$ is given by (6.129) as

$$F = \frac{1 - \Lambda}{\Lambda}\frac{\nu_E - p + 1}{\nu_H}, \tag{8.28}$$

   where $\Lambda$ is defined in (8.27), $\nu_E = N - k$, and $\nu_H = k - 1$. The partial $\Lambda$-statistic in (8.27) is distributed as $\Lambda_{1,\nu_H,\nu_E-p+1}$, and the partial $F$ in (8.28) is distributed as $F_{\nu_H,\nu_E-p+1}$.

3. Structure coefficients (correlations between variables and discriminant functions) - show how strongly each variable relates to each discriminant function.

4. Rotation - sometimes discriminant functions can be rotated to enhance interpretability, similar to factor analysis.

These methods help move beyond statistical significance to substantive understanding.

# 8 Scatter Plots

Visual inspection is crucial. Once discriminant scores are computed for individuals on the first one or two discriminant functions, scatter plots can be drawn.

- Two-group case: individuals can be plotted along the single discriminant axis.
- Several groups: plots of the first two discriminant functions provide a reduced-dimensional representation of group separation.

Scatter plots reveal:

- Which groups are most distinct,
- Overlap between groups,
- Outliers or misclassified observations.

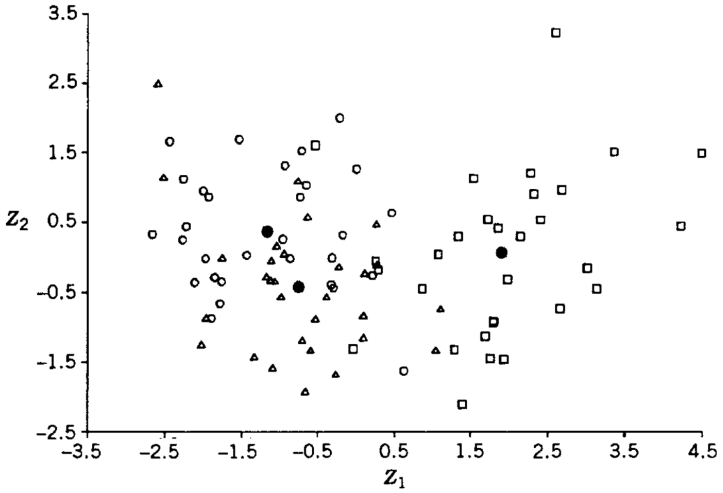


**Figure 8.4.** Scatter plot of discriminant function values for the football data of Table 8.3.

# 9 Stepwise Selection of Variables

In practice, not all variables contribute meaningfully to group separation. **Stepwise discriminant analysis** adds or removes variables according to their statistical contribution.

*Criteria for Entry/Removal*

- Wilks' Lambda: smaller values indicate stronger discrimination.
- Partial $F$-tests: assess significance of each variable when added to the model.

*Cautions*

- Stepwise methods can capitalize on chance and should be interpreted cautiously.
- Results may differ depending on order of entry and sample size.
- Best used as exploratory tools rather than definitive procedures.

**Example 8.9.** We use the football data of Table 8.3 to illustrate the stepwise procedure outlined in this section and in Section 6.11.2. At the first step, we carry out a univariate $F$ (using ordinary ANOVA) for each variable to determine which variable best separates the three groups by itself:

| Variable | $F$ | $p$-Value |
|---|---|---|
| WDIM | 2.550 | .0839 |
| CIRCUM | 6.231 | .0030 |
| FBEYE | 1.668 | .1947 |
| EYEHD | 58.162 | $1.11 \times 10^{-16}$ |
| EARHD | 22.427 | $1.40 \times 10^{-8}$ |
| JAW | 4.511 | .0137 |

Thus EYEHD is the first variable to "enter." The Wilks $\Lambda$ value equivalent to $F = 58.162$ is $\Lambda(y_1) = .4279$ (see Table 6.1 with $p = 1$). At the second step we calculate a partial $\Lambda$ and accompanying partial $F$ using (8.27) and (8.28):

$$\Lambda(y_r|y_1) = \frac{\Lambda(y_1, y_r)}{\Lambda(y_1)},$$

$$F = \frac{1 - \Lambda(y_r|y_1)}{\Lambda(y_r|y_1)} \frac{\nu_E - 1}{\nu_H},$$

where $y_1$ indicates the variable selected at step 1 (EYEHD) and $y_r$ represents each of the five variables to be examined at step 2. The results are

| Variable | Partial $\Lambda$ | Partial $F$ | $p$-Value |
|---|---|---|---|
| WDIM | .9355 | 2.964 | .0569 |
| CIRCUM | .9997 | .012 | .9881 |
| FBEYE | .9946 | .235 | .7911 |
| EARHD | .9525 | 2.143 | .1235 |
| JAW | .9540 | 2.072 | .1322 |

The variable WDIM would enter at this step, since it has the largest partial $F$. With a $p$-value of .0569, entering this variable may be questionable, but we will continue the procedure for illustrative purposes. We next check to see if EYEHD is still significant now that WDIM has entered. The partial $\Lambda$ and $F$ for EYEHD adjusted for WDIM

are $\Lambda = .424$ and $F = 58.47$. Thus EYEHD stays in. The overall Wilks' $\Lambda$ for EYEHD and WDIM is $\Lambda(y_1, y_2) = .4003$.

At step 3 we check each of the four remaining variables for possible entry using

$$\Lambda(y_r|y_1, y_2) = \frac{\Lambda(y_1, y_2, y_r)}{\Lambda(y_1, y_2)},$$
$$F = \frac{1 - \Lambda(y_r|y_1, y_2)}{\Lambda(y_r|y_1, y_2)} \frac{\nu_E - 2}{\nu_H},$$

where $y_1 =$ EYEHD, $y_2 =$ WDIM, and $y_r$ represents each of the other four variables. The results are

| Variable | Partial $\Lambda$ | Partial $F$ | $p$-Value |
|---|---|---|---|
| CIRCUM | .9774 | .981 | .3793 |
| FBEYE | .9748 | 1.098 | .3381 |
| EARHD | .9292 | 3.239 | .0441 |
| JAW | .8451 | 7.791 | .0008 |

The indicated variable for entry at this step is JAW. To determine whether one of the first two should be removed after JAW has entered, we calculate the partial $\Lambda$ and $F$ for each, adjusted for the other two:

| Variable | Partial $\Lambda$ | Partial $F$ | $p$-Value |
|---|---|---|---|
| WDIM | .8287 | 8.787 | .0003 |
| EYEHD | .4634 | 49.211 | $6.33 \times 10^{-15}$ |

Thus both previously entered variables remain in the model. The overall Wilks $\Lambda$ for EYEHD, WDIM, and JAW is $\Lambda(y_1, y_2, y_3) = .3383$.

At step 4 there are three candidate variables for entry. The partial $\Lambda$- and $F$-statistics are

$$\Lambda(y_r|y_1, y_2, y_3) = \frac{\Lambda(y_1, y_2, y_3, y_r)}{\Lambda(y_1, y_2, y_3)},$$
$$F = \frac{1 - \Lambda(y_r|y_1, y_2, y_3)}{\Lambda(y_r|y_1, y_2, y_3)} \frac{\nu_E - 3}{\nu_H},$$

where $y_1$, $y_2$, and $y_3$ are the three variables already entered and $y_r$ represents each of the other three remaining variables. The results are

| Variable | Partial $\Lambda$ | Partial $F$ | $p$-Value |
|---|---|---|---|
| CIRCUM | .9987 | .055 | .9462 |
| FBEYE | .9955 | .189 | .8282 |
| EARHD | .9080 | 4.257 | .0173 |

Hence EARHD enters at this step, and we check to see if any of the three previously entered variables has now become redundant. The partial $\Lambda$ and partial $F$ for each of these three are

| Variable | Partial $\Lambda$ | Partial $F$ | $p$-Value |
|---|---|---|---|
| WDIM | .7889 | 11.237 | $4.74 \times 10^{-15}$ |
| EYEHD | .6719 | 20.508 | $5.59 \times 10^{-8}$ |
| JAW | .8258 | 8.861 | .0003 |

Consequently, all three variables are retained. The overall Wilks' $\Lambda$ for all four variables is now $\Lambda(y_1, y_2, y_3, y_4) = .3072$.

At step 5, the partial $\Lambda$- and $F$-values are

| Variable | Partial $\Lambda$ | Partial $F$ | $p$-Value |
|---|---|---|---|
| CIRCUM | .9999 | .003 | .9971 |
| FBEYE | .9999 | .004 | .9965 |

Thus no more variables will enter.

We summarize the selection process as follows:

| Step | Variable Entered | Overall $\Lambda$ | Partial $\Lambda$ | Partial $F$ | $p$-Value |
|---|---|---|---|---|---|
| 1 | EYEHD | .4279 | .4279 | 58.162 | $1.11 \times 10^{-16}$ |
| 2 | WDIM | .4003 | .9355 | 2.964 | .0569 |
| 3 | JAW | .3383 | .8451 | 7.791 | .0008 |
| 4 | EARHD | .3072 | .9080 | 4.257 | .0173 |

□