Rabasová Marcela
Technical University of Ostrava, Czech Republic

Risk of morbidity in colorectal surgery

Keywords

colorectal surgery, morbidity, risk factors, laparoscopy, discriminant analysis

Abstract

This study examines the risk of morbidity for colorectal surgery undergoing patients. The main aim was to identify important risk factors that influence post-operative complications - morbidity, and to create a model to predict possible complications for a patient before surgery. The source data file contains information about 1177 patients who underwent colorectal surgery between 2001 and 2009 at the University Hospital Ostrava, Czech Republic. According to the surgeons' judgment the following seven independent variables were included in the analysis: Gender, BMI, American Society of Anaesthesiology (ASA) Classification, Stage of Disease, Number of Previous Operations, Surgical Technique and Operation Severity. Discriminant analysis was used for the data evaluation; statistical software SPSS 18 and NCSS 2004 were used for the calculations.

1. Introduction

Discriminant analysis is a multivariate statistical method, which can be used for the two main purposes. It may serve as a descriptive tool for describing differences among groups of units with regard to the vector of p random variables (descriptive discriminant analysis) or for predicting a group membership for a unit that has not been classified yet (predictive discriminant analysis). The model is built in the latter case, based on a set of observations for which the group memberships are known, and used to predict the appropriate class of a new observation with unknown group membership. Both these approaches are involved in the present study.

2. Discriminant analysis

The principles of discriminant analysis are introduced in [2]-[4]. Suppose that a multivariate random sample of the range n falls into H groups. If the groups of units can be demonstrated to differ on the level of p monitored quantitative variables, $X_1,...,$ X_p, which could be checked by an analysis of variance, the question can be put, to what degree these variables affect the unit membership.

2.1. Descriptive discriminant analysis

The basis of Fisher's concept of discriminant analysis is finding a linear combination $Y = \mathbf{b}^T\mathbf{x}$ of p monitored variables that would separate H groups better than any other such type with the intention that its within groups variance would be minimal and between groups variance would be maximal.

Using a notation H for the number of groups, n_h for the range of the h-th group, n for the total number of observations, $\mathbf{x}_{ih}$ for a vector of the values of variables $X_1,..., X_p$ on the i-th unit of the h-th group, $\bar{\mathbf{x}}$ for a vector of sample means and $\bar{\mathbf{x}}_h$ for a vector of sample means in the h-th group, the following formulas are valid:

- the total variance of the original p variables can be represented by the matrix **T**:

$$\mathbf{T} = \sum_{h=1}^{H}\sum_{i=1}^{n_h}(\mathbf{x}_{ih} - \bar{\mathbf{x}})(\mathbf{x}_{ih} - \bar{\mathbf{x}})^T ,$$

- the within groups variance by the matrix **E**:

$$\mathbf{E} = \sum_{h=1}^{H}\sum_{i=1}^{n_h}(\mathbf{x}_{ih} - \bar{\mathbf{x}}_h)(\mathbf{x}_{ih} - \bar{\mathbf{x}}_h)^T ,$$

- the between groups variance by the matrix **B**:

$$\mathbf{B} = \sum_{h=1}^{H}\sum_{i=1}^{n_h}(\bar{\mathbf{x}}_h - \bar{\mathbf{x}})(\bar{\mathbf{x}}_h - \bar{\mathbf{x}})^T .$$

Thus $\mathbf{T} = \mathbf{E} + \mathbf{B}$ and sum of squares $Q_B(Y)$ and $Q_E(Y)$, which represent the rate of between and within groups variance of a new variable Y, can be expressed as: $Q_B(Y) = \mathbf{b}^T\mathbf{B}\mathbf{b}$ and $Q_E(Y) = \mathbf{b}^T\mathbf{E}\mathbf{b}$. Minimal within groups variance and maximal between groups variance are managed if the ratio

$$F = \frac{Q_B(Y)}{Q_E(Y)} = \frac{\mathbf{b}^T\mathbf{B}\mathbf{b}}{\mathbf{b}^T\mathbf{E}\mathbf{b}}, \quad (1)$$

known as Fisher's discriminant criterion, has maximal value. Looking for the maximum of (1) the system of equations is received with the matrix equation

$$(\mathbf{B}\mathbf{E}^{-1} - \lambda\mathbf{I})\mathbf{b} = 0.$$

This system has a nontrivial solution provided that

$$\left|\mathbf{B}\mathbf{E}^{-1} - \lambda\mathbf{I}\right| = 0. \quad (2)$$

Characteristic equation (2) has r solutions, which are the eigenvalues $\lambda_1,\ldots, \lambda_r$ of the matrix $\mathbf{B}\mathbf{E}^{-1}$ ($\lambda_1 > \ldots > \lambda_r$). The eigenvector $\mathbf{b_1}$ associated with the largest eigenvalue λ_1 maximizes discriminant criterion (1).
Characteristic equation (2) does not determine the vector $\mathbf{b_1}$ uniquely, it determines only the proportions of its components. It is advisable to choose their concrete values so that the condition

$$\frac{1}{n-H}\mathbf{b_1}^T\mathbf{E}\mathbf{b_1} = 1 \quad (3)$$

was satisfied. Then criterion (1) could be expressed as

$$F = \frac{1}{n-H}\mathbf{b_1}^T\mathbf{B}\mathbf{b_1}$$

and the eigenvalue λ_1 represents the degree of between groups variance of the variable $Y_1 = \mathbf{b_1}^T\mathbf{x}$.
The linear combination $Y_1 = \mathbf{b_1}^T\mathbf{x}$ is called the first discriminant or also the first canonical variable. If the set of units, which is described by p variables, is divided into two groups, only one discriminant is sufficient for representing the total variance of the original variables.
If the measured values of the variables $X_1,\ldots, X_p$ on the i-th unit are substituted for $\mathbf{x}$, we obtain the so-called discriminant score of the unit. Using the constant

$$c_1 = -\sum_{k=1}^{p} b_{1k}\bar{x}_k = -\mathbf{b_1}^T\bar{\mathbf{x}}, \quad (4)$$

in the calculation ensures that the mean of the discriminant scores equals zero. Consequently, for the i-th unit, $i = 1,\ldots, n_h$, of the h-th group, $h = 1,\ldots, H$, the first discriminant score is computed as

$$y_{i1h} = c_1 + \sum_{k=1}^{p} b_{1k}x_{ihk}.$$

The computation of the vectors of average discriminant values in the groups is useful for obtaining the view on how the individual groups differ with regard to the canonical variable. These values are called the group centroids.
The coefficient b_{1k} indicates the individual impact of the k-th original variable X_k on the first canonical variable Y_1 provided that the other variables are constant. These coefficients are often standardized due to the better interpretation of results. Denoting $\mathbf{F}$ the diagonal matrix with the square roots of diagonal elements of matrix $\mathbf{E}$ the standardized coefficients are:

$$\mathbf{b_1^*} = \frac{1}{\sqrt{n-H}}\mathbf{F}\mathbf{b_1}. \quad (5)$$

The alternative approach to results' interpretation uses correlation coefficients between canonical variable and original variables (structure r's). A large absolute values of these coefficients indicate the importance of the original variables for respective discriminant. If the correlation coefficient is positive, larger values of the original variable lead to increasing the canonical variable value and vice versa. The vector of correlation coefficients for the first discriminant is given by the formula:

$$\mathbf{a_1} = \frac{1}{\sqrt{n-H}}\mathbf{F}^{-1}\mathbf{E}\mathbf{b_1}. \quad (6)$$

The significance of the canonical variables in discrimination can be tested by the Wilks'statistic

$$\Lambda = |\mathbf{E}| / |\mathbf{E}+\mathbf{B}|,$$

which has an F distribution in the event that $H = 2$. Otherwise the Bartlett's approximation can be used, where the quantity

$$V = c(-\ln\Lambda), \quad (7)$$

with $c = n - 1 - (p+H)/2$, has a χ^2 distribution

with $p(H\text{-}1)$ degrees of freedom.

2.2. Predictive discriminant analysis

Suppose that a random sample of n units falls into H groups. The p values of quantitative variables $X_1,...,$ X_p are at disposal for each unit as well as its group membership, which is represented by the value of an alternative or multivariate nominal variable called classifying criterion.
The goal of predictive discriminant analysis is to set up a rule, based on $n \times p$ data matrix that would predict group membership for an arbitrary population unit for which the values of the variables $X_1,..., X_p$ are known.
Let us assume a two-group classification problem where the distribution of the random vector $\mathbf{x}$ is a multivariate normal with the mean vectors $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ and the equal covariance matrices $\boldsymbol{\Sigma}$. The eigenvector $\mathbf{b}$ that maximizes the Fisher's discriminant criterion (1) may be expressed then as:

$$\mathbf{b} = k\,\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2),$$

where k is an arbitrary constant. Provided that condition (3) is satisfied this constant has a form

$$k = [\,(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^{\mathrm{T}}\,\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\,]^{-1/2}$$

and we can easily derive the following classification rule: Assign a unit represented by the vector of scores $\mathbf{x}$ to group 1 if

$$\mathbf{x}^{\mathrm{T}}\,\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) > 1/2(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^{\mathrm{T}}\,\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2),$$

otherwise to group 2. This rule corresponds to the idea of assigning a unit into that group which is the nearest in the sense of its distance from the group centroids. This classification rule leads to the confrontation of the values of the two functions:

$$LCF_1(\mathbf{x}) = \mathbf{x}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 - \frac{1}{2}\boldsymbol{\mu}_1^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1,$$

$$LCF_2(\mathbf{x}) = \mathbf{x}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2 - \frac{1}{2}\boldsymbol{\mu}_2^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2$$

which are called linear classification functions. Using these functions a unit is classified into the group with a higher *LCF* score. Such classification minimizes the total proportion of misclassification errors.
The parameter values $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ and $\boldsymbol{\Sigma}$ are seldom known in practice so their likelihood estimates are used in computations. However the total proportion of misclassified units is not minimal then.

2.3. Classification efficiency evaluation

The probability of the correct classification of units (hit rate) is an important information about the quality of used discriminant criterion. It could be assessed in several different ways.
A so-called resubstitution is one of the options, in which the discriminant criterion is used for classifying the same units as have served for its derivation. It is obvious that so gained assessment is overestimated.
Another option is to divide disposable data set into two parts. One part is used for discriminant criterion derivation, the second one for its verification. So gained assessment is unbiased but it requires a sufficiently large data set. Moreover the discriminant criterion is not as efficient as it could be in the case of inclusion all the data in the process of its creation.
A cross-validated method is also recommended (so called "jackknife" procedure), which subsequently creates the discriminant criterion using all units of the data set except the i-th, $i = 1,..., n$, which is then classified and the accuracy of this classification is checked. The assessment of the probability of the correct classification in this method is almost unbiased.

3. Study design

The source data set consisted of information about 1177 patients who underwent colorectal surgery between 2001 and 2009 at the University Hospital Ostrava, Czech Republic. The main aim of this study was to identify important risk factors that influence post-operative complications - morbidity, and to create a model to predict possible complications for a patient before surgery.
The seven outcome variables were selected on the basis of professional judgment of the surgeons: Gender, BMI, American Society of Anaesthesiology (ASA) Classification, Stage of Disease (SD), Number of Previous Operations (NPO), Surgical Technique (ST) and Operation Severity (OS). The variable Morbidity acts as a grouping variable.

3.1. Data conditioning

Numeric values were assigned to nominal variables in this way: Gender (1 – female, 0 – male), Stage of Disease (values 1 – 4 were assigned in accordance with histological tumour node metastasis (TNM) classification for colorectal cancer, value 0 represents benign disease), Surgical Technique (0 – open, 1 – laparoscopy), Operation Severity (values 2,

4, 8 were assigned in accordance with methods published by Copeland et al. [1]).
The two grouping-variable values were defined as 0 – patients without complications and 1 – patients with complications.
The total number of 90 missing values was found in the 1177 × 7 data matrix. All the 90 incomplete observations were excluded from the analysis.
Finally the 1087 × 7 data matrix was involved in our analysis with n_1 = 735 units in group 0 and n_2 = 352 units in group 1.

4. Analysis results

The data were analyzed using statistical software SPSS Version 18.0 and NCSS 2004.
Descriptive information for our seven outcome variables is given in *Table 1*.

Table 1. Descriptive statistics

Morbidity		Mean	Min	Max	Variance
0	ST	,53	0	1	,250
	Gender	,43	0	1	,245
	BMI	26,34	13,8	41,6	20,544
	ASA	2,22	1	4	,571
	NPO	,72	0	6	,850
	SD	2,21	0	4	2,016
	OS	4,35	2	8	2,730
1	ST	,50	0	1	,251
	Gender	,37	0	1	,234
	BMI	26,93	16,2	45,7	22,469
	ASA	2,40	1	4	,542
	NPO	,74	0	5	,786
	SD	2,05	0	4	2,089
	OS	4,85	2	8	3,805
Total	ST	,52	0	1	,250
	Gender	,41	0	1	,242
	BMI	26,53	13,8	45,7	21,224
	ASA	2,28	1	4	,568
	NPO	,72	0	6	,828
	SD	2,16	0	4	2,043
	OS	4,51	2	8	3,129

We assumed that the joint distribution of each of the seven outcome variables is approximately normal in each of the two groups. The Box test for covariance homogeneity provided no evidence that the population covariance matrices differ. The SPSS test results are presented in *Table 2*.

Table 2. Box test

Box's M	35,177
F Approx.	1,246
df1	28
df2	1795624,033
Sig.	,173

Because there is support for the equality of the two covariance matrices, we proceed with a multivariate analysis of variance (MANOVA). The hypothesis of equality of group means vectors was rejected at the significance level 1% (see *Table 3*), thus the influence of the seven monitored quantities upon morbidity was proved. The additional one-way tests for individual variables show significant differences among group means especially for the variables ASA Classification, Operation Severity and BMI. These three variables seem to have the greatest impact on possible morbidity. Discriminant analysis was used to confirm this fact.

Table 3. MANOVA

Term(DF) Test Statistic A(1): Morbidity	Test Value	F-Ratio	Prob. Level	(0,05)
Wilks' Lambda	0,958	6,73	0,000	Reject
Hotelling-Lawley Tr.	0,044	6,73	0,000	Reject
Pillai's Trace	0,042	6,73	0,000	Reject
Roy's Largest Root	0,044	6,73	0,000	Reject
ST	0,185	0,74	0,390	Accept
Gender	0,794	3,28	0,070	Accept
BMI	83,086	3,93	0,048	Reject
ASA	7,832	13,86	0,000	Reject
NPO	0,126	0,15	0,697	Accept
SD	6,283	3,08	0,079	Accept
OS	57,813	18,78	0,000	Reject

As the classification variable (morbidity) has only two values, the data set is divided into two groups and only one discriminant is satisfactory to represent the total variability of the seven initial variables. Its coefficients are shown in *Table 4*, the constant *c* given by the formula (4) is in the last row of the table.

Table 4. Canonical Discriminant Function Coefficients

	Function 1
ST	-,265
Gender	-,579
BMI	,034
ASA	,873
NPO	,072
SD	-,237
OS	,379
(Constant)	-3,753

An interpretation of the resulting group differences is based on correlation coefficients between each of the seven outcome variables and canonical discriminant function, which are given by the formula (6). These coefficients are reported in *Table 5* (Structure Matrix).

Table 5. Structure Matrix

	Function 1
OS	,632
ASA	,540
BMI	,289
Gender	-,264
SD	-,256
ST	-,125
NPO	,057

From the structure *r*'s we conclude that the variable with the greatest impact on morbidity is the variable Operation Severity with the coefficient 0.632, followed by the variables ASA Classification (0.540) and BMI (0.289). Larger values of these variables mean a larger value of the canonical variable and therefore, the greater risk of morbidity. (The group centroids are reported in *Table 6*). The variable that contributes the least to separation of patients with or without morbidity is the variable Number of Previous Operations (0.057).

Table 6. Group Centroids

Morbidity	Function 1
0	-,144
1	,301

The significance of the first canonical variable in discrimination was tested by the statistic given by the formula (7). The statistical test results reported in *Table 7* confirmed the importance of this variable (P = 0.000).

Table 7. Wilks' Lambda

Test of Function(s) dimension	Wilks' Lambda	Chi-square	df	P
0 1	,958	45,927	7	,000

The second aim of this study was to establish a prediction rule for predicting possible complications for a patient before surgery.
Because the Box test did not reject the hypothesis of the equality of the two covariance matrices, there is support for the use of a linear classification rule.
The possibility of deleting one or more predictors was sought at the beginning and the best results were obtained after deleting a variable Number of previous operations. Thus only six predictors were involved in the further computations: Gender, BMI, ASA Classification, Stage of Disease, Surgical Technique and Operation Severity.
The coefficients of the two linear classification functions were computed (see *Table 8*), which can serve for predicting morbidity for a patient before surgery. Given a set of six predictor scores for a new patient, a linear composite score for each group is found by multiplying each predictor score by the respective weight, summing these six products, and adding the constant. Each patient is then assigned to that group (with or without morbidity) for which the determined score is larger.

Table 8. Classification Function Coefficients

	Morbidity	
	0	1
ST	2,724	2,597
Gender	2,725	2,482
BMI	1,159	1,174
ASA	2,454	2,844
SD	1,140	1,031
OS	1,761	1,929
(Constant)	-25,084	-26,759

The classification efficiency evaluation is given in *Table 9*, often referred to as a confusion matrix.
The results of resubstitution are shown in the upper half of *Table 9* ("Original"). We can see that the total number of the correctly classified patients is 643 (440 without and 203 with morbidity), which is 59.2%. This number determines the probability of the correct classification. However this value is overestimated, which was mentioned in Chapter 2.3. So the cross-validated method was used to obtain an unbiased estimation of the probability of the correct classification. Its results are found in the bottom half of *Table 9* ("Cross-validated"). This method provides almost unbiased estimation of our

classification model accuracy, which is 58.3% ((437+197)/1087).

Table 9. Classification Results

		Mor-bidi-ty	Predicted Group Membership		Total
			0	1	
Original	Count	0	440	295	735
		1	149	203	352
	%	0	59,9	40,1	100,0
		1	42,3	57,7	100,0
Cross-validated	Count	0	437	298	735
		1	155	197	352
	%	0	59,5	40,5	100,0
		1	44,0	56,0	100,0

The risk of morbidity for patients undergoing colorectal surgery could be then predicted on the basis of the coefficients of the two linear classification functions reported in *Table 8* with the 58.3% probability. Regrettably, this number is so small that such a model cannot be used in practice. Different predictors have to be taken into account in future research.

5. Conclusion

Morbidity after colorectal operations depends on many factors. This study focused on the risk rate of the following seven factors: Gender, BMI, ASA Classification, Stage of Disease, Number of Previous Operations, Surgical Technique and Operation Severity. Discriminant analysis did not find the chosen input variables satisfactory enough to make a sufficient model for the prediction of morbidity, which means that a new choice of independent predictors is necessary. This task will be solved in the future.

The variable Operation Severity was marked as the variable with the greatest impact on possible morbidity, followed by the variables ASA Classification and BMI. Larger values of these variables mean greater risk of morbidity.

The other variables do not have such significant impact on morbidity, which is an important finding especially in the case of the variable Surgical Technique. This fact implies that there is no difference in morbidity for the two operation methods - laparoscopic and open.

Acknowledgments

This work is supported by The Ministry of Education, Youth and Sports of the Czech Republic. Project CQR 1M06047.

References

[1] Copeland, G.P., Jones, D. & Wakters, M. (1991). POSSUM: a scoring system for surgical audit. *Br. J. Surg*. 78, 356-360.

[2] Hebák, P. & Hustopecký, J. & Jarošová, E. & Pecáková, I. (2004). *Vícerozměrné statistické metody [1] (Multivariate statistical methods [1])*. Praha: Informatorium.

[3] Huberty, C.J. & Olejnik, S. (2006). *Applied MANOVA and Discriminant Analysis*. Wiley-Interscience.

[4] Neil, H.T. (2002). *Applied Multivariate Analysis*. Springer-Verlag, New York, USA.