Can we use categorical independent variable in discriminant analysis?

Asked 4 years, 8 months ago Active 3 years ago Viewed 14k times

14

In discriminant analysis, the dependent variable is categorical, but can I use a categorical variable (e.g residential status: rural, urban) along with some other continuous variable as independent variable in linear discriminant analysis?

edited Feb 22 '17 at 20:55

amoeba says Reinstate Monica
77.1k 22 238 286

asked Jun 26 '15 at 10:51

kuwoli
247 2 3 10

Similar question – ttnphns Jun 26 '15 at 12:43

2 Answers

14

Discriminant analysis assumes a multivariate normal distribution because what we usually consider to be predictors are really a multivariate dependent variable, and the grouping variable is considered to be a predictor. This means that categorical variables that are to be treated as predictors in the sense you wish are not handled well. This is one reason that many, including myself, consider discriminant analysis to have been made obsolete by logistic regression. Logistic regression makes no distributional assumptions of any kind, on either the left hand or the right hand side of the model.

Logistic regression is a direct probability model and doesn't require one to use Bayes' rule to convert results to probabilities as does discriminant analysis.

edited Jun 26 '15 at 11:44

Nick Cox
42.1k 6 95 138

answered Jun 26 '15 at 11:34

Frank Harrell
60.7k 4 121 261

Thank you Mr. Frank Harrell for your response. Actually i want to compare the results of discriminat analysis and logistic regression (logit model) using the same set of variable. So, for that purpose if i have to use the categorical variables in discriminant analysis as independent variable then is there any way? – kuwoli Jun 26 '15 at 11:44

6

The short answer is rather no than yes.

One preliminary note. It is difficult to say whether the variables which produce discriminant functions out of themselves should be called "independent" or "dependent". LDA is basically a specific case of Canonical correlation analysis, and therefore it is ambidirectional. It can be seen as MANOVA (with the class variable as the independent factor) or, when the class is dichotomous, as a linear regression of the class as the dependent variable. It is *not quite* legal therefore to always oppose LDA with one-directional regressions such as logistic one.

LDA assumes that the variables (those you called "independent") come from multivariate normal distribution, hence - all them continuous. This assumption is important for (1) classification stage of LDA and (2) testing significance of the discriminants produced at the extraction stage. The extracting of the discriminants itself does not need the assumption.

However LDA is quite robust to the violation of the assumption which is seen sometimes as a warranty to do it on **binary** data. In fact, some people do it. Canonical correlations (of which LDA is a specific case) can be done where both sets consist of binary or even dummy binary variables. Once again, there is no problem with the extraction of the latent functions; the problems with such application potentially arise when p-values or classifying objects are invoked.

From binary/ordinal variables one might compute tetrachoric/polychoric correlations and submit it to LDA (if the program allows to input correlation matrices in place of data); but then computation of discriminant scores on case level will be problematic.

A more flexible approach would be to turn categorical (ordinal, nominal) variables into continuous by *optimal scaling/quantification*. **Nonlinear canonical correlation analysis** (OVERALS). It will do it under the task to maximize canonical correlations between the two sides (the class variable and the categorical "predictors"). You may then try LDA with the transformed variables.

(Multinomial or binary) logistic regression may be another alternative to LDA.

edited Apr 13 '17 at 12:44

Community ♦
1

answered Jun 26 '15 at 12:24

ttnphns
44.1k ● 28 ● 194 ● 411

This is much more involved than just using a model that was intended for the situation (logistic regression). Discriminant analysis is not as robust as some think. It is easy to show with a single categorical predictor that is binary that the posterior probabilities form d.a. are not very accurate (e.g., predict the probability of an event given a subject's sex). – Frank Harrell Jun 26 '15 at 18:36