Project Assignment 2

Logistic Regression

A copy of the written report will be collected on the 17th of March.
The .R file should be e-mailed to Maria Durbán the same day.

1 Credit-scoring

If banks give a credit to a client, they are interested in estimating the risk that the client will not pay back the credit as agreed by contract. The aim of credit-scoring systems is to model or predict the probability that a client with certain risk factors is to be considered as a potential risk. Credit scoring methods became standard tool of banks and other financial institutions, direct marketing retailers and advertising companies to estimate whether an applicant for credit/goods will pay back his liabilities.

The success of credit scoring in credit cards issuing was a significant sign for the banks to use scoring methods to other products like personal loans, mortgage loans, small business loans etc. However, commercial lending is more heterogeneous, its documentation is not standardized within or across institutions and thus the results are not so clear. The growth of direct marketing has led to the use of scorecards to improve the response rate to advertising campaigns in the 1990s.

Most of the problems one must face when using credit scoring are rather technical than theoretical nature. First of all, one should think of the data necessary to implement the scoring. It should include as many relevant factors as possible. It is a trade-off between expensive data and between low accuracy due to not enough information. Banks collect the data from their internal sources (from the applicants previous credit history), from external sources (questionnaires, interviews with the applicants) and from third parties. From the applicants background the following information is usually collected: age, gender, marital status, nationality, education, number of children, job, income, lease rental charges, etc. The following questions from applicants credit history are especially interesting: "Has the applicant already a credit?", How much did he borrowed?, Has the applicant ever delayed his payment?, Does he ask for another credit as well? Under third parties we understand special houses oriented in collecting credit information about potential clients. The variables entering the credit scoring procedures should be chosen carefully, as the amount of the data may be vast indeed and thus computationally problematic.

2 The data set description

The data set consist of 1000 consumers' credit from a bank. The response variable of interest is "creditability", which is given in dichotomous form ($y = 0$ for creditworthy, $y = 1$ for not creditworthy). In addition, 20 covariates are assumed to influence creditability were collected.

2.0.1 Description of the variables

- Y **(response variable):** a factor with levels 'good', 'bad'. Payment of previous loans (good means "good payer").
- X_1 **(Account):** a factor with levels 'no running account', 'good running', 'bad running', quality of the credit clients bank account.
- X_2 **(Month):** a numeric vector, duration of loan in months.
- X_3 **(Ppag):** a factor with levels 'pre good payer' 'pre bad payer', if the client previously have been a good or bad payer.
- X_4 **(Use):** a factor with levels 'personal' 'professional', the intended use to which the loan is made.
- X_5 **(DM):** a numeric vector, the size of loan in monetary units.
- X_6 **(Sex):** a factor with levels 'woman' 'man', sex of the client.
- X_7 **(CivSt):** a factor with levels 'alone' 'not alone', civil state of the client.

The data set is divided into two samples of 500 individuals each, a training sample "training.txt" (used to estimate the model) a testing sample (used to validate the model) "testing.txt"

3 Assignment

1. Fit seven simple logistic regression model, one for each variable and comment the results (check whether the linear assumption is correct for the continuous covariates).
2. Fit a multivariate logistic regression model (take into account that the quality of the credit clients bank account and previous behavior of the client when paying loans might interact with other variables) and use the likelihood ratio test to determine which terms should be dropped from the model and compare the result with the model obtained using AIC criteria.
3. Based of this last model, obtain 90% confidence intervals for the odds ratios for loan use and civil state.
4. Interpret the coefficients of the model obtained in terms of odds-ratios.
5. Conduct the appropriate goodness of fit test, state the hypothesis, the decision rule and the conclusion.
6. Use the deviance residuals to check the adequacy of the model.
7. Use the H matrix to identify outlying X observations.
8. Do the appropriate plots for detecting influential observations.

9. To predict the credit status, you must identify the optimal cutoff. On the basis of the training sample, find the total error rate, the error rate for persons considered "good payers", and the error rate for persons considered "bad payers", for 100 different cutoff points between 0.1 and 0.9. Which of the cutoffs minimizes the total error rate? Looking at the area under the ROC curve, what do you conclude?

10. Use the prediction rule obtained before to predict the credit status of the other 500 individuals in the file "testing.txt". What are the total and the two component prediction error rates for the validation sample? How do these error rates compare with the ones obtained in the model fitting sample?

11. Combine the model fitting and validation sample and fit the model selected previously to the combined data. Are the estimated coefficients and the estimated standard errors similar to the ones obtained for the model fitting data set? Should they be? Comment.