Robust Regression With Both Continuous and Categorical Regressors in the Presence of Missing Values

Tsung-Chi Cheng
Department of Statistics
National Chengchi University
64 Chih-Nan Road, Section 2
Taipei 11623, Taiwan
chengt@nccu.edu.tw

Yu-Wen Wen
Department of Statistics
National Chengchi University
64 Chih-Nan Road, Section 2
Taipei 11623, Taiwan
g8354502@m0.nccu.edu.tw

1. Problems and approaches

Robust regression estimation and diagnostics have been widely discussed in the literature, where both response and regressors are usually continuous. It is very often that data are mixed with continuous and categorical regressors. Draper and Smith (1998) treat categorical regressors as the usual continuous ones using the least squares estimation, which is known to be very sensitive to outliers. Therefore it seems natural to extend the robust regression methods to the problem of this kind, such as M and S estimators. Hubert and Rousseeuw (1997) propose the RDL_1 estimator. They first downweights the leverage points in the space of the continuous regressors and then follows a weighted least absolute values fit as a function of continuous and categorical regressors. Maronna and Yohai (2000) consider two types of estimates for the problems. The first one is also a weighted L_1 estimate. The other consists of an M estimate for the coefficients of categorical regressors and an S estimate for those continuous ones because the former would not be robust enough and the later would be too expensive for computation when using either one in the model. Their simulation shows that the weighted L_1 estimate would be better when the dimension of continuous regressors is smaller, whereas the M-S estimate is better when it is equal or greater than 4, especially for high contamination.

Little and Schluchter (1985) analyze the incomplete data with continuous and categorical variables. The swamping and masking effects caused by outliers in incomplete data when using the EM algorithm (Dempster *et al.* 1977) in both continuous response and regressors have been shown by Atkinson and Cheng (2000). We here combine the EM with the M-S estimates to

identify potential outliers and also to impute missing values. The robust procedure proceeds as follows:

1. Take a random sample of size $s = p + q + 1$ from the original data.
2. Impute the missing values by EM to obtain the complete data and the estimates of interest based on the subset of s cases.
3. Use M-S estimate to fit the complete data.

The procedure will be repeated several times to seek the optimum solution of the M-S estimates. We also extend the forward search for the least trimmed squares regression of Atkinson and Cheng (2000) to the problem.

2. An illustrative example: ozone concentration data

The data were collected in State College, Pennsylvania, from July 8 through August 27, 1991. There are nine independent variables including eight continuous variables and a categorical variable indicating whether the child stayed near the home for the whole day. The response variable is 12-hour average daytime personal ozone concentration. There are 89 cases in the data where missing values are present in both continuous and categorical variables. The details about these data can be seen in Liu and Wypij (1994). We apply our algorithms to these data, which lead to different conclusions from the previous results.

REFERENCES

Atkinson, A.C. and Cheng, T.-C. (2000). On robust linear regression with incomplete data. *Comp. Statist. Data Anal.* **33**, 361-380.

Dempster, A.P., Laird, M. and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Roy. Statist. Soc. Ser.* B, **39**, 1-38.

Draper, N.R. and Smith, H. (1998). *Applied Regression Analysis*, 3rd ed., New York: John Wiley.

Hubert, M. and Rousseeuw, P.J. (1997). Robust regression with both continuous and binary regressors. *J. Statist. Plan. Infer.*, **57**, 153-163.

Little, R.J.A. and Schluchter, M.D. (1985). Maximum likelihood estimation for mixed continuous and categorical data with missing values. *Biometrika*, **72**, 497-512.

Maronna, R.A. and Yohai, V.J. (2000). Robust regression with both continuous and categorical predictors. *J. Statist. Plan. Infer.*, **89**, 197-214.

Wypij, D. and Liu, L.-J.S. (1994) Prediction models for personal ozone exposure assessment. In *Case Studies in Biometry*, edited by N. Lange, *et al.*, New York: John Wiley, pp. 41-56.