

# Mixtures of Logit Regressions Detection with Forward Search

## *Stima di modelli di regressione logistica con distribuzioni miste mediante l'approccio Forward Search*

M. Bini, V. Del Pantà, M. Velucchi

**Abstract** The forward search is a general method to detect multiple outliers and to determine their effect on inference about models fitted to data. From the monitoring of a series of statistics based on subsets of data of increasing size we obtain multiple views of any hidden structure. Sometimes, some features emerge unexpectedly during the progression of the forward search only when a specific combination of forward plots is inspected at the same time. In this paper, we use a set of new robust graphical tools on a mixture of logit regressions. We use simulated data and we show the dynamic interaction with different “robust plots” to highlight the presence of groups of outliers and regression mixtures in the context of logit regression and highlight the effect that these hidden groups provide on the fitted model.

**Abstract** *La Forward Search è un approccio di analisi innovativo robusto che parte da un insieme di dati ridotto, privo di unità anomale ed include sequenzialmente le rimanenti osservazioni in base ad una misura via via crescente di “anomalia” delle stesse. Attraverso una combinazione efficace di modellazione statistica e di grafici diagnostici, la forward search costituisce uno strumento potente che individua la presenza di valori anomali (singoli o raggruppati), e permette di valutare il loro effetto sui risultati delle analisi tradizionali. In questo studio proponiamo l'uso di tale approccio nella stima di un modello di regressione logistica dove dati simulati provengono da popolazioni binomiali differenti. Grazie all'analisi grafica sarà possibile identificare unità o gruppi anomali che solo apparentemente appartengono alla stessa popolazione.*

**Key words:** Forward Search, Logit Regression, Robustness

---

Matilde Bini, Vieri Del Pantà, Margherita Velucchi  
Università Europea di Roma, Via degli Aldobrandeschi 190, 00163 Roma. e-mail: mbini@unier.it  
(corresponding author), vieri.delpanta@unier.it, margherita.velucchi@unier.it

## 1 Introduction

The forward search is a general approach for the detection of outliers and hidden structure in data and to build robust models. The basic idea of the forward search is simple: starting from small subsets of data, observations that are close to the fitted model are iteratively added to the observations used in parameter estimation. As the subset grows, parameter estimates are monitored, while test statistics and measures of fit such as residuals are registered and investigated. The first contribution on this topic refers to the work of [1], but the power and the number of applications (both univariate and multivariate) of this procedures considerably increased during the last decade thanks to the use of diagnostic monitoring [2]. One of the useful characteristics of the forward search is that the number of data trimming is not fixed in advance, but is chosen conditionally on the dataset. Increasing size subsets from the bulk of data are fitted in sequence and a whole series of subsets can be easily explored. As the subset size increases, the method of fitting moves from very robust to highly efficient maximum likelihood methods providing a good compromise between robustness and statistical efficiency [8]. In addition, sometimes during the progression of the forward search we can detect the existence of mixtures of distributions and the presence of outliers which have to be taken into account in the analysis. These features have to be investigated together in order to give an exhaustive description of a complex problem. In this paper, following [7] and [8], we adapt and develop a robust graphical tools set on a mixture of logit regressions.

We use simulated data and we show the dynamic interaction with different "robust plots" to highlight the presence of groups of outliers and regression mixtures, as well as to show the effects that these hidden groups produce on the fitted model.

## 2 The Model: Mixtures of Logit Regressions with the Forward Search

The forward search is based on a robust fit on very few observations [2]. Then, a successive fit is done with larger subsets. The initial subset is identified using the least median of squares method [9] that guarantees that no outliers are included in the initial subset. Formally [2, p. 31]: let  $\mathbf{Z} = (X, y)$  a data matrix of dimension  $n \times (p + 1)$ . If  $n$  is moderate and  $p \ll n$  the choice of the initial subset can be performed by exhaustive enumeration of all  $\binom{n}{p}$  distinct  $p$ -tuple  $S_{i_1, \dots, i_p}^{(p)} \equiv \{z_{i_1}, \dots, z_{i_p}\}$ , where  $z_{i_j}^T$  is the  $i_j$ -th row of  $\mathbf{Z}$ , for  $1 \leq i_1, \dots, i_p \leq n$  and  $i_j \neq i_{j'}$ .

Specifically, let  $\iota^T[i_1, \dots, i_p]$  and let  $e_{i, S_i^{(p)}}$  be the least squares residual for the unit  $i$  given the model has been fitted with the observations in  $S_i^{(p)}$ . The initial subset is  $S_*^{(p)}$  which satisfies

$$e_{[med], S_*^{(p)}}^2 = \min_i [e_{[med], S_i^{(p)}}^2] \quad (1)$$

where  $e_{[k], S_i^{(p)}}^2$  is the  $k$ -th ordered squared residual among  $e_{i, S_i^{(p)}}^2$ , with  $i = 1, \dots, n$  and  $med = \text{integer part of } (n + p + 1)/2$ . If  $\binom{n}{p}$  is too large, the choice is made using 3000  $p$ -tuples sampled from  $\mathbf{Z}$  matrix.

In this procedure, the subset size is increased by one and the model refitted to the observations with the smallest residuals for the increased subset size. The initial subset  $S_*^{(m)}$  of dimension  $m \geq p$  is increased by one and the new subset  $S_*^{(m+1)}$  consists of  $m + 1$  units with the smallest ordered residuals  $e_{[k], S_*^{(m)}}^2$ . The model is refitted to the new subset and the procedure continues increasing subset sizes until all the data are fitted, i.e. when  $S_*^{(m)} = S^{(n)}$ . The result is an ordering of the observations by closeness to the assumed model. Usually one observation enters the subset used for fitting, but sometimes two or more observations enter the subset as one or more leave. The result is an ordering of the observations by closeness to the assumed model.

The model we fit at each step  $m$  of the forward search is a logit regression having the general form

$$\ln\left(\frac{\hat{\pi}_{im}}{1 - \hat{\pi}_{im}}\right) = x_i^T \hat{\beta}_m = \hat{\beta}_{1m} + \hat{\beta}_{2m}x_{i2} + \dots + \hat{\beta}_{pm}x_{ip} \quad (2)$$

where  $i = 1, \dots, n$ ,  $m = p + 1, p + 2, \dots, n$ ,  $x_i^T = 1, x_{i2}, \dots, x_{ip}$  with  $p$  denoting the number of covariates, and  $\hat{\pi}_{im}$  is the expected response for the  $i$ -th observation at step  $m$ .

The estimated probability for the observation  $i$  at step  $m$  is

$$\hat{\pi}_{im} = \frac{\exp(x_i^T \hat{\beta}_m)}{1 + \exp(x_i^T \hat{\beta}_m)}. \quad (3)$$

Residuals are the typical measure of the discrepancy of a regression model and they are generally defined as the difference between what is observed ( $y_i$ ) and what is predicted by the model ( $\hat{\pi}_{im}$ ). As regards generalized linear models, residuals based on the deviance function are used to measure this discrepancy [5]. The deviance residual ( $r_{Dim}$ ) of the  $i$ -th observation at step  $m$  is defined as

$$r_{Dim} = \text{sign}(y_i - \hat{\pi}_{im}) \sqrt{\max(0, d_{im})}, \quad (4)$$

where  $d_{im}$  is the  $i$ -th component (at step  $m$ ) of the deviance function,  $D_m = \sum r_{Dim}^2$ , defined for a logit model as follows

$$D_m = 2 \sum \left\{ y_i \log\left(\frac{y_i}{\hat{\pi}_{im}}\right) + (1 - y_i) \log\left(\frac{1 - y_i}{1 - \hat{\pi}_{im}}\right) \right\}. \quad (5)$$

During the forward search, all relevant statistics may be monitored (estimators, t-stats, residuals etc.). One of the most important plots shows residuals at each step of

the forward search. Large values of the residuals or sudden changes in the value of the residual sum of squares show the existence of outliers or clusters of observations coming from different distributions (mixtures).

Some authors [8] found good approximations to the forward graphs of the minimum deletion residual and suggested a procedure for automatic outlier detection during the forward search. Sometimes, subsets of units or events occurring at specific steps of the process emerge during the progression of the forward search. Some authors [6] instead of inspecting them one by one, suggest to use two plots simultaneously, possibly in combination with other plots and by interacting with the graphics. Often, it is crucial to find the right trade-off between detail and scope of the data region inspected, by means of zoom and unzoom tools. In this paper, we extend this procedure to logit regression.

### 3 The data set

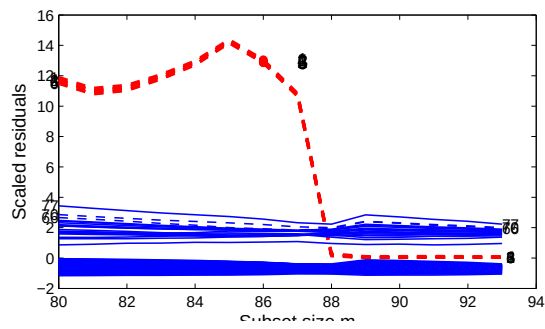
To show the power of the forward search procedure in detecting possible mixtures in the data that may affect results of estimation, a simulation study has been performed. As previously mentioned, the study is focused on the case of binomial data.

Firstly, a data set of 93 units with a binary response from a binomial distribution with  $\pi = 0.2$  and three covariates (one categorical and two continuous) has been generated, as an example for the estimation of a robust logit regression based on the forward search. Secondly, to evaluate the trajectories of the minimum deletion deviance residuals obtained from fitting a robust logit regression, empirical envelopes for 1%, 50% and 99% confidence intervals have been simulated using this data set. Finally, in order to highlight the presence of outliers, the data matrix has been modified, by shocking the values of the first six units of the data set (both the response and covariates).

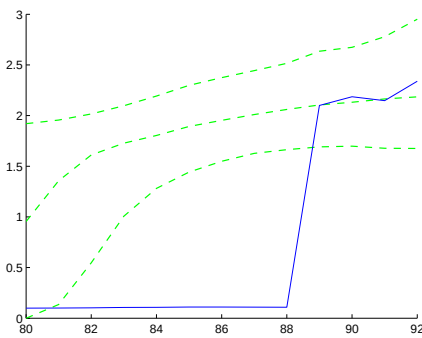
The results obtained applying the forward search in the estimation of the logit regression show the presence of a group of units different from the bulk of data.

The monitoring of scaled deviance residuals reported in Figure 1 reveals the presence of 6 trajectories having a shape higher than those of the remaining units of data set. They are isolated units and they do not enter exactly at the end of the procedure. Similarly, it is possible to see the effect they also produce looking at the minimum deletion residual plot, as shown in Figure 2. From steps 80 to 88 the line shows a peak out of the lower 1% band and intersects again the 50% line when  $m = 89$ , due to the their effect.

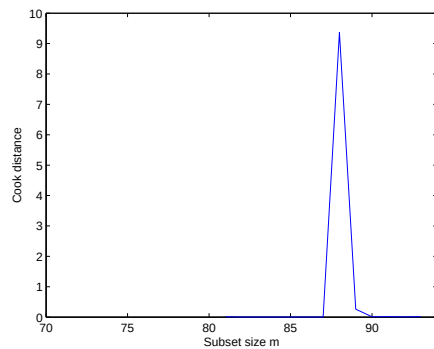
The influential importance of the six observations is shown by the plot of the Cook statistic in Figure 3. The interpretation of the peak at  $m = 88$  is that the addition of one observation of this group during the procedure causes a significant change in the parameter value (it springs up very high). Then, values of the statistic after this step are small because the introduction of the remaining observations do not modify the parameter values.



**Fig. 1** Forward plot of scaled residuals



**Fig. 2** Minimum deletion residual outside subset with 1%, 50% and 99% confidence envelopes



**Fig. 3** The effect of including observations 1-6 showed by Cooks distance

## References

1. Atkinson, A.C.: Fast very robust methods for the detection of multiple outliers. *Journal of the American Statistical Association* **89**, 1329–1339 (1994)

2. Atkinson, A.C., Riani, M.: Robust diagnostic regression analysis. Springer, New York (2000)
3. Atkinson, A.C., Riani, M., Cerioli, A.: Exploring multivariate data with the forward search. Springer, New York (2004)
4. Atkinson, A.C., Riani, M.: Building regression models with the forward search. *Journal of Computing and Information Technology*. **15**, 1330–1336 (2008)
5. McCullagh, P., Nelder, J.A.: Generalized linear models. Chapman and Hall, London (1989)
6. Perrotta, D., Riani, M., Torti, F.: New robust dynamic plots for regression mixture detection. *Advances in Data Analysis and Classification* (2009) doi: 10.1007/s11634-009-0050-y
7. Riani, M., Cerioli, A., Atkinson, A. C., Perrotta, D., Torti, F.: Fitting Mixtures of Regression Lines with the Forward Search. F. Fogelman-Souli et al. (Eds.) IOS Press (2008)
8. Riani, M., Atkinson, A.C., Cerioli, A.: Finding an unknown number of multivariate outliers. *Journal of the Royal Statistical Society, Series B*, **71**, 447–466 (2009)
9. Rousseeuw, P.J.: Least median of squares regression. *Journal of the American Statistical Association* **79**, 871–880 (1984)