

# The Effect of Link Function on the Monitoring of Logistic Regression Profiles

Mehdi Koosha, Amirhossein Amiri

**Abstract**— In some real case problems, a relationship between a response variable and one or more explanatory variables called as profile is desirable to be monitored over time instead of the response variable itself. There are many techniques in the literature for monitoring the profiles. A special case of the profiles is where the response variable follows a binomial distribution known as Logistic regression profile. In the logistic regression profile, a link function relates the mean of response variable to explanatory variables. There are many kinds of link function in the literature of logistic regression methods such as Logit, Probit, Log-Log, and so on. This paper investigates the effect of applying various link functions on the performance of  $T^2$  control chart in monitoring the parameters of a logistic profile.

**Index Terms:** Logistic regression, Profile monitoring,  $T^2$  control chart, Link function

## I. INTRODUCTION

There are many situations in which monitoring the relationship between response variable and explanatory variables called as profile is desirable instead of only monitoring the quality characteristics. Different types of profiles such as simple linear profile, multiple linear regression and polynomial profile, and so on are investigated in the literature of profiles monitoring. Many methods have been proposed for monitoring different types of profiles by researchers. For example; see [1], [2] and [3] for monitoring simple linear profiles and [4] for polynomial profiles monitoring.

In most of these researches, the response variable is assumed to be normal but this assumption is violated in some situations. In the area of monitoring non-normal profiles, [5] have proposed 5  $T^2$  based methods to monitor logistic profiles in which the response variable is binomial.

In logistic regression profiles, a link function is applied to make a linear relationship between the explanatory variables and the expected value of the response variable [6, 7]. This paper considers the effect of using various link functions including Logit, Probit, Log-Log and Comp Log-Log on the parameter estimation of logistic profiles as well as on the performance of  $T^2$  control chart which is used to monitor logistic profiles. The rest of this paper is classified as follows:

Manuscript received March 09, 2011; revised April 02, 2011.

Mehdi Koosha is with Industrial Engineering Department, Shahed University, Tehran, Iran. (Corresponding author; phone: (+9821) 51212065; e-mail: koosha@shahed.ac.ir).

Amirhossein Amiri is with Industrial Engineering Department, Shahed University, Tehran, Iran. (e-mail: amiri@shahed.ac.ir).

Section 2 describes the algorithm of logistic profile parameter estimation. Section 3 introduces various kinds of link functions. Section 4 explains the  $T^2$  control chart and its application in profile monitoring. Section 5 evaluates the effect of different link functions on the parameter estimation of logistic profile and on the performance of  $T^2$  control chart. Conclusions and future researches have been proposed in the last section.

## II. PARAMETER ESTIMATION OF A LOGISTIC PROFILE

When the observations are assumed to be normal, Least Square Error (LSE) method can be applied to estimate the parameters of a profile. But in the case that the response variable follows a Bernoulli distribution, this method leads to unbiased estimators which do not have minimum variance. So the maximum likelihood estimation (MLE) method can be used in this case. An iterative algorithm should be used to estimate the parameters in logistic regression profiles by using MLE method [7].

Suppose that there are  $m$  independent profiles ( $i=1,2,\dots,m$ ) and the observations in each level of the explanatory variable ( $j=1,2,\dots,p$ ) are repeated for  $n$  times.

For estimating of parameters in a logistic profile, [5] applied an approach that is explained below:

Start the algorithm with an initial  $\hat{\beta}$  which could be obtained by any method such as LSE.

**Step 1:** Set  $l=0$ . Determine the value of  $X$  (Vector of explanatory variables) and observe the response variables in each level.

**Step 2:** Determine a vector of  $\pi_i^{(l)}$  that contains elements  $\pi_{ij}^{(l)}$  which shows the probability of success in an experiment in  $j^{\text{th}}$  level of  $X$  in  $i^{\text{th}}$  profile in  $l^{\text{th}}$  iteration.

**Step 3:** Set  $l=l+1$ . Calculate  $\eta_i^{(l)} = X_i^T \hat{\beta}^{(l)}$  for any iteration of the algorithm.

**Step 4:** Update the values of  $\pi_i^{(l)}$  with Eq.1

$$\pi_i = g[E(Y_i)] = X_i' \beta, \quad (1)$$

where  $g$  is the inverse link function which is used to relate the values of  $\eta_i^{(l)} = X_i' \beta$  and the expected value of each observation ( $\pi_i^{(l)}$ ). Description about link functions and inverse link functions is given in section 3.

**Step 5:** Compute  $W$  and  $\mu$  by Eq.2 and Eq.3, respectively.  $\mu$  is the vector of expected values for response variables.

$$W_i^{(l)} = \text{diag}[n\pi_{ij}^{(l)}(1-\pi_{ij}^{(l)})] \quad (2)$$

$$\mu_i^{(l)} = [n\pi_{ij}^{(l)}] \quad (3)$$

**Step 6:** Compute  $q_i^{(l)}$  by applying Eq.4.

$$q_i^{(l)} = \eta_i^{(l)} + (W_i^{(l)})^{-1} (Y_i - \mu^{(l)}) \quad (4)$$

$Y_i$  is the vector of observations in  $i^{\text{th}}$  profile.

**Step 7:** Update the estimation for  $\beta$  by Eq.5.

$$\hat{\beta}^{(l+1)} = (X^T W_i^{(l)} X)^{-1} X^T W_i^{(l)} q_i^{(l)} \quad (5)$$

If  $\hat{\beta}^{(l+1)} - \hat{\beta}^{(l)} < \varepsilon$ , then  $\hat{\beta}^{(l+1)}$  can be used as the final estimation of  $\beta$ .

Else set  $l=l+1$  and go to step 3

### III. LINK FUNCTION

Link function ( $g$ ) is used to make a relationship between the mean and the explanatory variables.

$$g[E(Y_i)] = X_i' \beta \quad (6)$$

There are many kinds of link functions in the literature some of them are summarized in the Table 1. [1]

Table 1- Link functions

Link Function	Link Function	Inverse Link Function
	$\eta = g(\mu)$	$\mu = g^{-1}(\eta)$
Log-Log	$-\ln\{-\ln(\mu)\}$	$\exp\{-\exp(-\eta)\}$
Logit	$\ln\left(\frac{\mu}{1-\mu}\right)$	$\frac{e^\eta}{1+e^\eta}$
Probit	$\phi^{-1}(\mu)$	$\phi(\eta)$
Com loglog	$-\ln\{-\ln(1-\mu)\}$	$1 - \exp\{-\exp(-\eta)\}$

If we apply any link function in a logistic regression parameter estimation procedure, the results will be unbiased. In each iterations, 30 binomial profiles are generated and the number of observations in each level of the explanatory variable ( $X$ ) is equal to 30. Besides, we have 9 levels for  $X$  including  $[\log(0.1) \log(0.2) \log(0.3) \log(0.4) \log(0.5) \log(0.6) \log(0.7) \log(0.8) \log(0.9)]$ . The expected value of response variable is  $X_i' \beta = \beta_0 + \beta_1 x = 3 + 2x$ . The number of iterations is 5000. The mean and standard deviation of the estimations for  $\beta_0$  and  $\beta_1$  are summarized in Table 2.

Table 2. - Mean of the estimations for coefficients of the profile under various link functions

Link function	Logit	Probit	Log-Log	Comp log-log
Parameter Estimation's Mean	3.05,2.04	3.07,2.05	3.14,2.09	3.13,2.05
Parameter Estimation's standard deviation	0.34,0.28	0.34,0.32	0.48,0.36	0.48,0.38

As shown in Table 2, the estimations are unbiased and the deviations from their nominal values (3, 2) are due to simulation error. However, the standard deviations of the estimations are changed under various link functions. As shown in Table 2, Logit and comploglog link functions leads to minimum and maximum variances of regression parameters' estimators, respectively. Section 5 shows that change of the link function has destructive effect on the performance of the  $T^2$  control chart. This effect is due to the inflation in the standard deviation of the estimators.

### IV. $T^2$ HOTELING CONTROL CHART

For monitoring Logistic profiles, [5] proposed 5 T-squared statistics including  $T^2$  based on sample mean and covariance matrix ( $T_h^2$ ),  $T^2$  based on sample average and moving ranges ( $T_r^2$ ),  $T^2$  based on sample average and intra profile pooling ( $T_i^2$ ),  $T^2$  based on Minimum Volume Ellipsoid ( $T_{MVE}^2$ ) and  $T^2$  based on Minimum Covariance Determinant ( $T_{MCD}^2$ ). This paper uses  $T_i^2$  because as shown by [5] this statistic outperforms the others in detecting either small or large shifts. This statistic is defined in Eq.7 as follows:

$$T_i^2 = (\hat{\beta}_t - \bar{\beta})^T S^{-1} (\hat{\beta}_t - \bar{\beta}), \quad (7)$$

where  $\hat{\beta}_t$  is the estimated vector of regression parameters for  $t^{\text{th}}$  profile. The  $\hat{\beta}_t$  vector supposed to follow a  $p$ -variate normal distribution with mean  $\beta$  and variance-covariance ( $X^T W X$ )<sup>-1</sup>.  $\bar{\beta}$  and  $S$  are the mean vector and covariance matrix of the parameters estimators and computed by Eq.8 and Eq.9, respectively.

$$\bar{\beta} = \frac{\sum_{t=1}^k \beta_t}{k} \quad (8)$$

$$S = \frac{1}{k} \sum_{t=1}^k \text{var}(\hat{\beta}_t) = \frac{1}{k} \sum_{t=1}^k (X^T W_t X)^{-1} \quad (9)$$

$W_t$  is defined in Eq.2. The UCL for the  $T^2$  control chart in Eq.7 is usually determined by simulation to obtain a predetermined value for probability of Type I error.

### V. THE EFFECT OF LINK FUNCTIONS ON THE PERFORMANCE OF $T_i^2$ CONTROL CHART

In this section, a simulation study has been done in order to illustrate the effect of applying various link functions on the performance of the  $T_i^2$  control chart in monitoring of logistic regression profile. The results are shown in Table 3. In this simulation study,  $X_t$  vector of controllable variables is  $[\log(0.1) \log(0.2) \log(0.3) \log(0.4) \log(0.5) \log(0.6) \log(0.7) \log(0.8) \log(0.9)]$ . Number of profiles generated ( $m$ ) and number of observations ( $n$ ) in each level of the explanatory variable ( $X$ ) is 30. The linear predictor is  $\eta_i = 3 + 2X_i$ . The upper control limits for  $T_i^2$  control charts under different link functions are given in Table 3. These control limits are set by simulation to obtain the probability of Type I error equals to 0.05 for all control charts. The power of each control chart in detecting different step shifts is obtained by simulation and the results are summarized in Table 3. Note that shifts in Table I are in the scale of  $\sigma$ . For example (1,2) means that  $\beta_{0,out} = \beta_{0,in} + 1\sigma_0$  and  $\beta_{1,out} = \beta_{1,in} + 2\sigma_1$  in which  $\sigma_0$  and  $\sigma_1$  are computed by Eq. 10.

$$S_0 = \begin{pmatrix} \sigma_0^2 & \rho\sigma_0\sigma_1 \\ \rho\sigma_0\sigma_1 & \sigma_1^2 \end{pmatrix} = (X^T W X)^{-1}, \quad (10)$$

where  $\rho$  is the correlation coefficient between logistic regression parameters estimators.

Table 3-Probability of signal for various shifts under different link functions

Link function \ Shift	Logit	Log-Log	Comploglog	Probit
	UCL=15.079	UCL=33.5	UCL=33.3	UCL=16.8
0,0	0.0546	0.0506	0.0518	0.0518
1,1	0.0914	0.0752	0.0816	0.1314
1,0	0.1658	0.1342	0.1498	0.1446
1,2	0.249	0.0836	0.0886	0.1904
2,3	0.4372	0.1522	0.1028	0.3870
2,0	0.7036	0.4082	0.1434	0.5486
1,3	0.7140	0.1246	0.1298	0.4370
2.5,0	0.9198	0.5316	0.5540	0.8194

The results in Table 3 are also depicted in Fig. 2. As shown in Fig. 2, the performance of  $T_i^2$  control chart has been changed by applying different link functions. We can also conclude that the best link function for logistic regression is Logit.

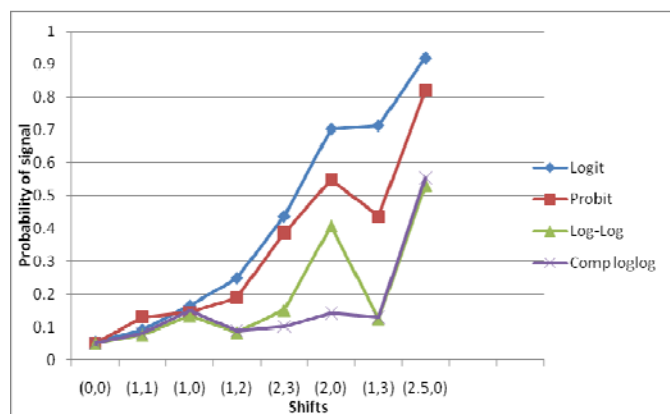


Fig 1 - Comparison between the performances of  $T_i^2$  control charts by applying various link functions

## VI. CONCLUSIONS

In this paper we showed the effect of using different link functions on the performance of  $T_i^2$  control chart which has been suggested by [5] for future research. The studies showed that Logit link function leads to the best performance for  $T_i^2$  control charts in detecting either small or large shifts. We have just investigated the effect of link function on the performance of one of the  $T^2$  control charts proposed by [5]. As a future research, one can investigate other control charts and check which control chart is more robust to various link function.

## REFERENCES

- [1] Kang, L., Albin, S. L., On-Line monitoring when the process yields a linear profile, *Journal of Quality Technology*, 2000, 32, 418-426.
- [2] Kim, K., Mahmoud, M.A., Woodall, W.H., On the monitoring of linear profiles, *Journal of Quality Technology*, 2003, 35, 317-328.
- [3] Mahmoud, M.A., Parker, P.A., Woodall, W.H., Hawkins, D.M., A change point method for linear profile data, *Quality and Reliability Engineering International*, 2007, 23, 247-268.
- [4] Kazemzadeh, R. B., Noorossana, R., Amiri, A., Phase I monitoring of polynomial profiles, *Communications in Statistics-Theory and Methods*, 2008, 37, 1671-1686.
- [5] Yeh, A. B., Huwang, L., Li, Y. M., Profile monitoring for a binary response, *IIE Transactions*, 2009, 41, 931-939.
- [6] Kleinbaum, D., Klein, M., *Logistic regression a self-learning Text*, Springer, 2002.
- [7] Agresti, A., *An introduction to categorical data analysis*, Wiley, 2007.