行政院國家科學委員會專題研究計畫 成果報告

異質變異數矩陣之穩健估計
研究成果報告(精簡版)

計 畫 類 別 ：個別型
計 畫 編 號 ：NSC 98-2118-M-004-006-
執 行 期 間 ：98 年 08 月 01 日至 99 年 08 月 31 日
執 行 單 位 ：國立政治大學統計學系

計 畫 主 持 人：鄭宗記

計畫參與人員：此計畫無其他參與人員

報 告 附 件 ：出席國際會議研究心得報告及發表論文

處 理 方 式 ：本計畫可公開查詢

中 華 民 國 99 年 09 月 26 日

On Robust Estimation of the Heteroscedasticity Covariance Matrix

Tsung-Chi Cheng*

1 Introduction

In a normal regression model, the assumption of homogenous variance is not always appropriate. When the disturbance process in a regression model entails heteroscedasticity, the standard inference procedure becomes invalid because of the inappropriate estimation of the standard error. A conventional way to overcome this problem in statistical and econometric literature is to specify the model with an assumed error structure and apply maximum likelihood estimation, generalized squares, or other approaches, such as residual maximum likelihood (Verbyla 1993). In addition, Bianco, Boente, and di Rienzo (2000) and Hallina and Mizera (2001) consider the robust estimator when outliers exit in a heteroscedastic regression model. However, these approaches require a specific function of the error variance. There is usually little or no guidance regarding the form of heteroscedasticity though.

Robust estimations and diagnostics for linear regression models with the assumption of constant errors have been widely discussed (see Atkinson (1985); Rousseeuw and Leroy (1987); Atkinson and Riani (2000)). Swamping (i.e. when inliers appear as outlying) and masking (i.e. when outliers appear as inlying) effects due to multiple outliers can be avoided by robust diagnostics. Both outliers and heteroscedasticity in the data also can lead to the inflation of the estimate of scale and deteriorate both the swamping and masking effects. For a successful analysis with regard to outliers and leverage points, a robust estimation is required, preferably one with a high

*Department of Statistics, National Chengchi University, 64 ZhihNan Road, Section 2, Taipei 11623, Taiwan; E-mail: chengt@nccu.edu.tw; TEL:+886 2 29393091#81132; FAX:+886 2 29398024

breakdown point. The (finite) sample breakdown point of an estimator is the smallest proportion of observations that, when altered, can cause the value of the estimator to become arbitrarily large or small. Therefore, one of the desirable properties for a robust estimator is a high breakdown point that can handle multiple outliers.

The approach proposed in this article employs the weighted least absolute deviation (WLAD) estimator suggested by Hubert and Rousseeuw (1997) and Giloni, Simonoff, and Sengupta (2006) to deal simultaneously with outliers and heteroscedasticity in the linear regression model without specifying the variance function. The difficulty is differentiating those observations that inflate the variation and belong to outlying points from those attributable to the (natural) heteroscedastic structure of the data. A jigsaw plot that uses the simulated envelopes of Atkinson (1985) for the absolute standardized residuals can represent both characteristics for each case in the dataset. Furthermore, plugging the resulting residuals into the estimation of the heteroscedasticity consistent covariance matrix (HCCM) yields a robust quasi-t test for the estimated coefficients.

2 Weighted least absolute deviation estimator

Consider the linear model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{1}$$

where $\boldsymbol{y}$ is an $n \times 1$ vector of the response variable, $\boldsymbol{\beta} = (\beta_0, \cdots, \beta_p)$ denotes $(p+1) \times 1$ regression coefficients, $\boldsymbol{X} = (\boldsymbol{x}_1^T, \cdots, \boldsymbol{x}_n^T)$ is an $n \times (p+1)$ design matrix, and $\boldsymbol{\epsilon}$ is the $n \times 1$ vector of random errors. The random vector $\boldsymbol{\epsilon} = \{\epsilon_1, \cdots, \epsilon_n\}$ is assumed to be independent and follows $MN(\mathbf{0}, \boldsymbol{\Omega})$, where $\boldsymbol{\Omega} = \text{diag}(\omega_1, \cdots, \omega_n)$. Given the form of the variance function, ω_i, the maximum likelihood estimation (MLE) for model (1) has been discussed by Harvey (1976) and Aitkin (1987) and the residual maximum likelihood (REML) estimation is presented by Patterson and Thompson (1971), Harville (1974), and Cooper and Thompson (1977). However, outliers can influence both MLE and REML (see Cheng (2010)). Without specifying the variance function, the WLAD procedure for model (1) essentially follows both Hubert and Rousseeuw (1997) and Giloni *et al.* (2006).

Hubert and Rousseeuw (1997) propose the RDL_1 estimator for robust regression with both continuous and categorical predictors. The RDL_1 consists of three stages: identifying leverage points, downweighting the leverage points when estimating the parameters, and estimating the residual scale. To adapt the RDL_1 estimator for model (1) with heteroscedastic errors, this study first computes the robust distance of continuous regressors (if discrete regressors are included in $\boldsymbol{X}$, they are excluded at this stage, as in Hubert and Rousseeuw (1997)), as follows:

$$RD(\boldsymbol{x}_i) = \sqrt{(\boldsymbol{x}_i - \boldsymbol{t})^T \boldsymbol{C}^{-1} (\boldsymbol{x}_i - \boldsymbol{t})}, \quad i = 1, \cdots, n, \tag{2}$$

where $\boldsymbol{t}$ and $\boldsymbol{C}$ are the robust location and scale estimates of the $\boldsymbol{X}$ matrix, respectively. Hubert and Rousseeuw (1997) use the minimum volume ellipsoid (MVE) estimator for $\boldsymbol{t}$ and $\boldsymbol{C}$. These distances (2) can identify the leverage points for the space of continuous regressors and serve as the weights for estimating the regression coefficients by a weighted L_1 procedure in the second stage. The current approach applies the minimum covariance determinant (MCD) estimator to obtain the robust location and scale estimates of the $\boldsymbol{X}$ matrix, and then obtains the distance (2) for the weights. Both MVE and MCD estimators provide a high breakdown of the robust estimation of multivariate location and shape (Rousseeuw and Leroy 1987). Moreover, Butler, Davies, and Jhun (1993) show that the MCD estimator has better theoretical properties than the MVE. Woodruff and Rocke's (1994) empirical results show that the MCD is preferable to the MVE for their applications. Croux and Haesbroeck (1999) discuss other statistical properties and the robustness of the MCD.

At the second stage, the parameters $\boldsymbol{\beta}$ of model (1) can be estimated by

$$\min_{\beta} \sum_{i=1}^{n} w_i |r_i(\boldsymbol{\beta})|, \tag{3}$$

where $r_i(\boldsymbol{\beta}) = y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}$, and

$$w_i = \min \left\{ 1, \frac{p}{(RD(\boldsymbol{x}_i))^2} \right\} \tag{4}$$

for $i = 1, 2, \cdots, n$. The final stage requires calculating the estimate of the scale of the residuals:

$$\hat{\sigma} = 1.4826 \text{median}_i |r_i|, \tag{5}$$

where the constant 1.4826 leads to a consistent estimator under a normality assumption. The standardized residual for the ith case then can be defined as

$$t_i = \frac{r_i}{\hat{\sigma}}. \tag{6}$$

An observation is flagged as an outlier if its absolute value from equation (6) exceeds 2.5 with an assumption of constant errors. The breakdown property of RDL_1 is referred to Hubert and Rousseeuw (1997).

Maronna and Yohai (2000) suggest there must be some null residuals by a well-known property of the weighted L_1 estimate (3) lead to an underestimation of the error variability. Instead of equation (5), they suggest using

$$\hat{\sigma}^* = 1.4826 s^*, \tag{7}$$

where s^* is the median of absolute non-null residuals. The standardized residual (6) then is replaced by $\hat{\sigma}^*$, as follows:

$$t_i^* = \frac{r_i}{\hat{\sigma}^*}. \tag{8}$$

Maronna and Yohai's suggestions are appropriate for the following discussion. Furthermore, the entire computation is easy to conduct. Both MCD and L_1 estimations are built-in functions in R and other commercial statistical packages, such as SPLUS and SAS.

Giloni *et al.* (2006) discuss some properties of the weighed L_1 estimator and suggest using the weight $\sqrt{\min_j(h_{jj})h_{ii}}$, where h_{ij} is the (i,j)th element of the hat matrix, $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$. This method enables them to match the results of Ellis and Morgenthaler (1992), who argue that breakdown is related to distance rather than squared distance.

2.1 Jigsaw plot

As mentioned in the previous section, the critical values that flag the possible outliers for standardized residuals (6) and (8) are ± 2.5 under constant errors for model (1). However, these values may not fulfil the requirement when heteroscedasticity exists in the data, and therefore, the estimation uses the unequal weights. To identify outlying

cases with an approach based on the WLAD estimate, the proposed method uses the half normal plot with envelopes (see Section 4.2 of Atkinson (1985)). The envelope then determines the threshold for the identification of outliers.

To construct the envelopes, the matrix $\boldsymbol{X}$ is fixed, such that the weight (4) remains the same throughout the simulation procedure. The response variable is generated from a normal distribution with zero mean, and its variance appears in Cook and Weisberg's (1983) study as follows for the ith case:

$$\sigma^2\{(1-h_{ii})w_i + \sum_{k\neq i} w_i h_{ik}/(1-h_{ii})\}. \tag{9}$$

Suppose that on the mth simulation of the n observations, the absolute values of the standardization residual is denoted by t^*_{mi}, $i = 1, \cdots, n$. The corresponding order is given by $t^*_{m(i)}$. The simulation can be repeated a fixed number of times, such as 80 times, which roughly coincides with the previous cut point, 2.5, for the percentile under normality. The simulated limits are given by

$$\begin{aligned} t^*_{l(i)} &= \min_m t^*_{m(i)}, \\ t^*_{u(i)} &= \max_m t^*_{m(i)}, \end{aligned} \tag{10}$$

where $t^*_{l(i)}$ and $t^*_{l(i)}$ form the lower and upper envelopes, respectively. The lower bound may be (near) 0 for all i due to null residuals of the L_1 estimate. This scenario results in a jigsaw shape when plotting the lower and upper bounds from equation (10) together.

There are two kind of estimates for σ in equation (9), which differentiate outliers and heteroscedastic structures in the data. One is the estimated scale from equation (7), and the other is the standardized residual in equation (8) for case i. The former provide a threshold for the test of outliers, whereas the latter reflects the heteroscedastic error for each case. The weighted hat matrix also might be used instead of $\boldsymbol{H}$, which may yield some different jigsaw plots according to the data structure. Nevertheless, the conclusion does not vary in either case. The weighted hat matrix is denoted by $\boldsymbol{H}_w = \boldsymbol{X}_w(\boldsymbol{X}_w^T\boldsymbol{X}_w)^{-1}\boldsymbol{X}_w^T$, where $\boldsymbol{X}_w = \boldsymbol{W}\boldsymbol{X}$, and $\boldsymbol{W}$ is a matrix with w_i for the ith diagonal element and zero's for all off-diagonal elements.

2.2 Types of outlying cases

Rousseeuw and van Zomeren (1990) propose a diagnostic plot of Studentised residuals (based on the least median squares estimate) versus robust distances (using MVE) of the $\boldsymbol{X}$ matrix, on the basis of which they classify the different types of outliers. Rousseeuw and Van Driessen (1999) adapt it as a D-D plot for the standardized residuals (using the least trimmed squares estimates) versus robust distances based on MCD. Employing a similar idea, this subsection describes the data types for the linear regression model with heteroscedastic error.

Part (a) of Figure 1 presents a scatter plot of 30 simulated data points, in which they are classified according to regular point, good leverage point, vertical outlier, and bad leverage point. Applying the proposed approach to this dataset results in Part (b) of Figure 1, which shows the diagnostic plot based on the WLAD estimate and locates all observations into their corresponding areas. The cutoff lines to separate these areas are ± 2.5 and $\sqrt{\chi^2_{p,0.975}}$ (here, $p = 1$) for horizontal and vertical lines, respectively.

In Figure 1, Parts (c) and (d) are the jigsaw plots of the absolute values of the standardized residuals (8), denoted by the symbol $\times$, together with the envelopes generated by using (7) and (8) for the σ of equation (9), respectively. The former reveals that cases 29 and 30 are outlying cases, whereas the latter indicates which observations are attributable to the heteroscedastic structure in the data. Both plots coincide with the data pattern in Part (a) in terms of outliers and heteroscedasticity. The jigsaw shape in Part (c) provides the corresponding threshold for each observation by identifying whether it is an outlier, which takes into account the unequal weight property. Regular points with labeled case numbers yield larger values for the upper bound of the envelope in Part (d), which indicates the source of heteroscedasticity.

2.3 Artificial data

This subsection identifies high leverage points for the heteroscedastic regression model using simulated data. The following model is employed for good data

$$y_i = 1 + 2x_i + \epsilon_i \ , \qquad i = 1, \ldots, 38, \tag{11}$$

where x_i is generated from a uniform distribution $U(1,7)$ and $\epsilon_i = \sqrt{5x_i}\eta$. Here, $\eta \overset{iid}{\sim} N(0, 0.5^2)$. For bad data, the following bivariate normal distribution applies:

$$\begin{pmatrix} y_i \\ x_i \end{pmatrix} \sim MN\left(\begin{pmatrix} 3 \\ 20 \end{pmatrix}, \begin{pmatrix} 0.5 & 0 \\ 0 & 0.5 \end{pmatrix}\right), \quad i = 39, \ldots, 50.$$

Therefore, the data contain 24% outliers. Part (a) of Figure 2 shows the scatter plot for this artificial dataset, which actually resulted from an adaption of the famous synthetic data provided by Rousseeuw (1984).

The analysis of this data set begins by applying the MM estimator of Yohai (1987). The function lmrob in the library robustbase provides the solution of an MM-regression estimator. It uses a bi-square re-descending score function, and by default, it returns a highly robust and highly efficient estimator. The computation of the robust standard errors relies on the formulas provided by Croux, Dhaene, and Hoorelbeke (2003). Part (b) in Figure 2 presents the resulting diagnostic plot, which identifies 12 bad leverage cases as good leverage points and three good observations as vertical outliers. Without taking into account heteroscedastic errors, the swamping and masking effects exist even though the high breakdown estimator is used.

On the contrary, WLAD successfully identifies leverage points, as shown in the standardized residual plot and jigsaw plot of Parts (c) and (d) of Figure 2, respectively. The latter presents the heteroscedastic configuration as a dashed line as well as the cutoff values for outliers, depicted by a doted line. These leverage points cause heteroscedasticity in the data as well, which may partly explain the existence of the masking and swamping effects when the MM estimate gets applied to these data.

3 Estimation of heteroscedasticity consistent covariance matrix

White (1980) proposes an estimator of the variance covariance matrix of the least squares regression coefficient that is consistent in certain conditions, which is also known as the HC0 estimator. Tests based on a heteroscedasticity-consistent covariance matrix (HCCM) estimator are popular in application, because there is no need to specify the structural form of heteroscedasticity, and it is easy to compute.

Despite this popularity of White's HC0 estimator, several critiques and improvements have been proposed. Chesher and Jewitt (1987) show that the estimator exhibits bias even for large samples under certain regression designs. MacKinnon and White (1985) thus propose a close variant form of HC0 based on the unreplicated "almost unbiased estimator" of Horn , Horn, and Duncan (1975). Long and Ervin (2000) compare several HCCM estimators with Monte Carlo studies. Cribari-Neto (2004) proposes a new estimator for HCCM that takes into account the leverage effect of the design matrix on associated quasi-t tests. For some additional alternatives, readers should turn to Bera , Surprayitno. and Premaratne (2002).

Several authors also argue that the leverage points are more decisive for the finite sample behavior than the degree of heteroscedasticity in the HCCM estimation (see Chesher and Jewitt 1987; Kempthorne and Mendel 1990; Furno 1997; Cribari-Neto and Zarkos 2001; and Cribari-Neto 2004). However, these approaches are based on ordinary least squares (OLS) estimator, which is notoriously influenced by outliers. Zhou and Portnoy (1998) provide inferential results derived from heteroscedastic models based on regression quantiles, where the median regression is a special case. The variance function for weights is specified but may be estimated by regressing the local estimates of standard errors on regression.

The most popular regression estimator for model (1) is the OLS estimate of $\boldsymbol{\beta}$, as given by $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$. It is unbiased, and the corresponding variance covariance matrix $Var(\hat{\boldsymbol{\beta}}) = \boldsymbol{\Psi}$ denoted by

$$\boldsymbol{\Psi} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\Omega}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}. \tag{12}$$

Under homoscedasticity, this equation simplifies to $(\boldsymbol{X}^T\boldsymbol{X})^{-1}\hat{\sigma}^2$, where $\hat{\sigma}^2 = \sum e_i^2/(n-p)$ denotes the estimated variance of model (1). The OLS estimator for the regression coefficients is consistent but inefficient in the general linear model with heteroscedastic errors.

Various heteroscedasticity consistent (HC) estimators for $\boldsymbol{\Psi}$ have been suggested and can be constructed by plugging an estimate of $\hat{\boldsymbol{\Omega}} = \text{diag}(\hat{\omega}_1, \cdots, \hat{\omega}_n)$ into equation (12). These estimators differ in the choice of $\hat{\omega}_i$, given as follows:

$$\text{constant:} \;\; \hat{\omega}_i \;\; = \;\; \hat{\sigma}^2, \tag{13}$$

$$\text{HC0: } \hat{\omega}_i = e_i^2, \tag{14}$$

$$\text{HC1: } \hat{\omega}_i = \frac{n}{n-p}\hat{e}_i^2, \tag{15}$$

$$\text{HC2: } \hat{\omega}_i = \frac{e_i^2}{1-h_{ii}}, \tag{16}$$

$$\text{HC3: } \hat{\omega}_i = \frac{e_i^2}{(1-h_{ii})^2}, \tag{17}$$

$$\text{HC4: } \hat{\omega}_i = \frac{e_i^2}{(1-h_{ii})^{\delta_i}}, \tag{18}$$

where $\delta_i = \min\{4, h_{ii}/\bar{h}\}$, and $\bar{h}$ is the average of $h_{ii}, i = 1, \cdots, n$.

Equation (13) yields the standard estimator $\hat{\mathbf{\Psi}}$ for homoscedastic errors; the other all lead to different kinds of HC estimators. The estimators HC1, HC2, and HC3, according to MacKinnon and White (1985), improve the performance in small samples. Long and Ervin (2000) conduct an extensive simulation study based on sample samples and conclude that HC3 provides the best performance. Cribari-Neto (2004) instead recommends the estimator HC4, which takes into account the leverage effect of the design matrix. Zeileis (2004) provides an R package for all these HC estimates.

3.1 Robust HCCM estimator

It is well-established that HCCM is influenced by outliers, especially the leverage points (e.g., Cribari-Neto (2004)). All the estimates in equations (13) to (18), based on OLS, provide consistent results under heteroscedasticity, but this property may vanish when outliers exist in the data. This subsection therefore contains the robust estimation for (12) that avoids the influence of outliers, using the approach discussed in Section 2.

With the WLAD estimate, it is possible to consider the choices of $\hat{\omega}_i$, which are analogous to those of classical HCCM as follows:

$$\text{RHC0: } \hat{\omega}_i = r_i^2, \tag{19}$$

$$\text{RHC1: } \hat{\omega}_i = \frac{n}{n-p-1}r_i^2, \tag{20}$$

$$\text{RHC2: } \hat{\omega}_i = \frac{r_i^2}{1-h_{wii}}, \tag{21}$$

$$\text{RHC3: } \hat{\omega}_i = \frac{r_i^2}{(1-h_{wii})^2}. \tag{22}$$

In these cases, r_i is the ith residual based on the weighted L_1 estimate in equation (3), and h_{wii} denotes the ith diagonal element of the weighted hat matrix $\boldsymbol{H}_w$.

3.2 Simulation study

A examination of the capability of the proposed robust test procedure simulates model (1) under heteroscedasticity. The data generation follows the method described in Subsection 2.3, with a focus on the problem of bad leverage points. The good data can be generated by the following model:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_4 x_4 + \epsilon,$$

where the parameter $\beta_0 = 20$, and all other β are 1. The values of x_1 can be generated from a uniform distribution with values between 0 and 10, and the other x variables have values between 0 and 20. To simplify the study, only one explanatory variable, x_1, is related to the error function, namely, $\epsilon \sim N(0, \exp(\sigma_i^2))$, where $\sigma_i = \delta_0 + \delta_1 x_{1i}$. The values of the parameters are set to $\delta_0 = 0.001$ and $\delta_1 = 0.0$, 0.3, 0.6, or 0.9 (these different values of δ_1 are denoted by data types a, b, c, and d, respectively, in the subsequent discussion). The differing values of δ_1 produce a constant error and a relatively moderate to very severe degree of heteroscedasticity. The "bad" data derive from a multivariate distribution with the following form:

$$\begin{pmatrix} \boldsymbol{x}_i^T \\ y_i \end{pmatrix} \sim MN\left(\begin{pmatrix} \boldsymbol{x}_m^T \\ y_m - 20 \end{pmatrix}, diag(0.25^2, \cdots, 0.25^2)\right),$$

where x_m is the maximum value of good x_2 plus 10, and y_m denotes the smallest values of good y_i. The adjustment of x_m and y_m allows for more distinct distances between bad and good data when heteroscedasticity is more severe in the data.

The sample sizes are 50, 100, 200, and 400, and each dataset contains 0%, 5%, 10%, 15%, and 20% outliers. One thousand replications compare the coverage of $\boldsymbol{\beta}$ when the robust results from equations (19) to (22) help estimate equation (12). The comparison of the ratio of the number of tests $H_0 : \beta_j = \beta_{0j}$, $j = 0, 1, \cdots, 4$, successfully rejects the null hypothesis in 1000 simulated data sets. Figure 3 shows

the average of the empirical p-values for tests of β_j according to different approaches, proportions of outliers, sample sizes, and the severity of the heteroscedasticity. All classical HCCM using (14) to (18) are spoiled by outliers, though they may have good properties without any single outlier in the data, as discussed in Long and Ervin (2000). Therefore, only the results pertaining to HC3 are reported here.

The robust standard error of the MM estimator proposed by Croux *et al.* (2003) provides reasonable results when the degree of heteroscedasticity is not severe and/or the proportions of outliers are not too large in the data. The sample size is a factor that influences the behavior of the test with regard to data types c and d with the MM estimate. The different versions of the robust HCCM from equations (19) to (22) supply similar results, regardless of the configurations of the data structure. All empirical p-values are close to 0.05. These four robust HCCM yield almost the same results for the same dataset, whereas the classical HCCM in equations (14) to (18) produce quite varied outcomes. The quasi-t tests using the robust HCCM also can resist bad leverage points.

References

Aitkin, M., 1987. Modelling variance heterogeneity in normal regression using GLIM. Applied Statistics 36, 332-339.

Atkinson, A.C., 1985. Plots, Transformations and Regression. Oxford University Press, Oxford.

Atkinson, A.C., Riani, M., 2000. Robust Diagnostic and Regression Analysis. Springer, New York.

Bianco, A., Boente, G., di Rienzo, J., 2000. Some results for robust GM-based estimators in heteroscedastic regression models. Journal of Statistical Planning and Inference 89, 215-242.

Bera, A.K., Surprayitno, T., Premaratne, G., 2002. On some heteroskedasticity-robust estimators of variance matrix of the least-squares estimators. Journal of

Statistical Planning and Inference 108, 121-136.

Butler, R.W., Davies, P.L., Jhun, M., 1993. Asymptotics for the minimum covariance determinant estimator. The Annals of Statistics 21, 1385-1400.

Cheng, T.-C., 2010. Discussion: the forward search: theory and data analysis. Journal of the Korean Statistical Society 39, 153-159.

Chesher, A., Jewitt, I., 1987. The bias of a heteroskedasticity consistent covariance matrix estimator. Econometrica 55, 1217-1222.

Cook, R. D., 1986. Assessment of local influence (with discussion). Journal of the Royal Statistical Society B 48, 133-169.

Cook, R.D., Weisberg, S., 1983. Diagnostics for heteroscedasticity in regression. Biometrika 70, 1-10.

Cooper, D.M., Thompson, R., 1977. A note on the estimation of the parameters of the autoregressive-moving average process. Biometrika 64, 625-628.

Cribari-Neto, F., 2004. Asymptotic inference under heteroskedasticity of unknown form. Computational Statistical and Data Analysis 45, 215-233.

Cribari-Neto, F., Zarkos, S.G., 2001. Heteroskedasticity-consistent covariance matrix estimation: White's estimator and the bootstrap. Journal of Statistical Computation and Simulation 68, 391-411.

Cribari-Neto, F., Zarkos, S.G., 2004. Leverage-adjusted heteroskedastic bootstrap methods. Journal of Statistical Computation and Simulation 74, 215-232.

Croux, C., Dhaene, G., Hoorelbeke, D., 2003. Robust standard errors for robust estimators. Discussion Papers Series 03.16, K. U. Leuven, CES.

Croux, C., Haesbroeck, G., 1999. Influence function and efficiency of the minimum covariance determinant scatter matrix estimator. Journal of Multivariate Analysis 71, 161-190.

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

Ellis, S.P., Morgenthaler, S., 1992. Leverage and breakdown in L_1-regression. Journal of the American Statistical Association 87, 143-148.

Furno, M., 1997. A robust heteroskedasticity consistent covariance matrix estimator. Statistics 30, 201-219.

Giloni, A., Simonoff, J.S., Sengupta, B., 2006. Robust weighted LAD regression. Computational Statistical and Data Analysis 50, 3124-3140.

Greene, W., 1997. Econometric Methods, 3rd ed., Prentice-Hall, Englewood Cliffs, NJ.

Hallina, M., Mizera, I., 2001. Sample heterogeneity and M-estimation. Journal of Statistical Planning and Inference 93, 139-160.

Harvey, A.C., 1976. Estimating regression models with multiplicative heteroscedasticity. Econometrika 38, 375-386.

Harville, D.A., 1974. Bayesian inference for variance components using only error contrasts. Biometrika 61, 383-385.

Horn, S.D., Horn R.A., Duncan, D.B., 1975. Estimating heteroscedastic variances in linear model. Journal of the American Statistical Association 70, 380-385.

Hubert, M., Rousseeuw, P.J., 1997. Robust regression with both continuous and binary regressors. Journal of Statistical Planning and Inference 57, 153-163.

Kempthorne, P.J., Mendel, M.B., 1990. Comment, Journal of the American Statistical Association 85, 647-648.

Koenker, R., Bassett, G., 1978. Regression quantiles. Econometrica 46, 211-244.

Long, J.S., Ervin, L.H., 2000. Using heteroskedasticity consistent standard errors in the linear regression model. American Statistician 54, 217-224.

MacKinnon, J.G., White, H., 1985. Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. Journal of Econometrics 29, 305-325.

Maronna, R.A., Yohai, V.J., 2000. Robust regression with both continuous and categorical predictors. Journal of Statistical Planning and Inference 89, 197-214.

Patterson, H.D., Thompson, R., 1971. Recovery of inter-block information when block sizes are unequal. Biometrika 54, 545-554.

Rousseeuw, P.J., 1984. Least median of squares regression. Journal of the American Statistical Association 79, 871-880.

Rousseeuw, P.J., Leroy, A.M., 1987. Robust Regression and Outlier Detection. John Wiley, New York.

Rousseeuw, P.J., van Driessen, K., 1999. A fast algorithm for the minimum covariance determinant estimator. Technometrics 41, 212-223.

Rousseeuw, P.J., van Zomeren, B.C., 1990. Unmasking multivariate outliers and leverage points (with discussion). Journal of the American Statistical Association 85, 633-651.

Verbyla, A.P., 1993. Modelling variance heterogeneity: residual maximum likelihood and diagnostics. Journal of the Royal Statistical Society B 55, 493-508.

Wei, B.-C., Hickernell, F.J., 1996. Regression transformation diagnostics for explanatory variables. Statistica Sinica 6, 433-454

Weisberg, S., 1980. Applied Linear Regression. John Wiley, New York.

White, H., 1980. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. Econometrica 48, 817-838.

Woodruff, D.L., Rocke, D.M., 1994. Computable robust estimation of multivariate location and shape in high dimension using compound estimators. Journal of the American Statistical Association 89, 888-896.

Yohai, V.J., 1987. High breakdown-point and high efficiency estimates for regression. The Annals of Statistics 15, 642-665.

Zeileis, A., 2004. Econometric computing with HC and HAC covariance matrix estimators. Journal of Statistical Software 11, 1-17.

Zhou, K.Q., Portnoy, S.L., 1998. Statistical inference on heteroscedastic models based on regression quantiles. Journal of Nonparametric Statistics 10, 239-260.

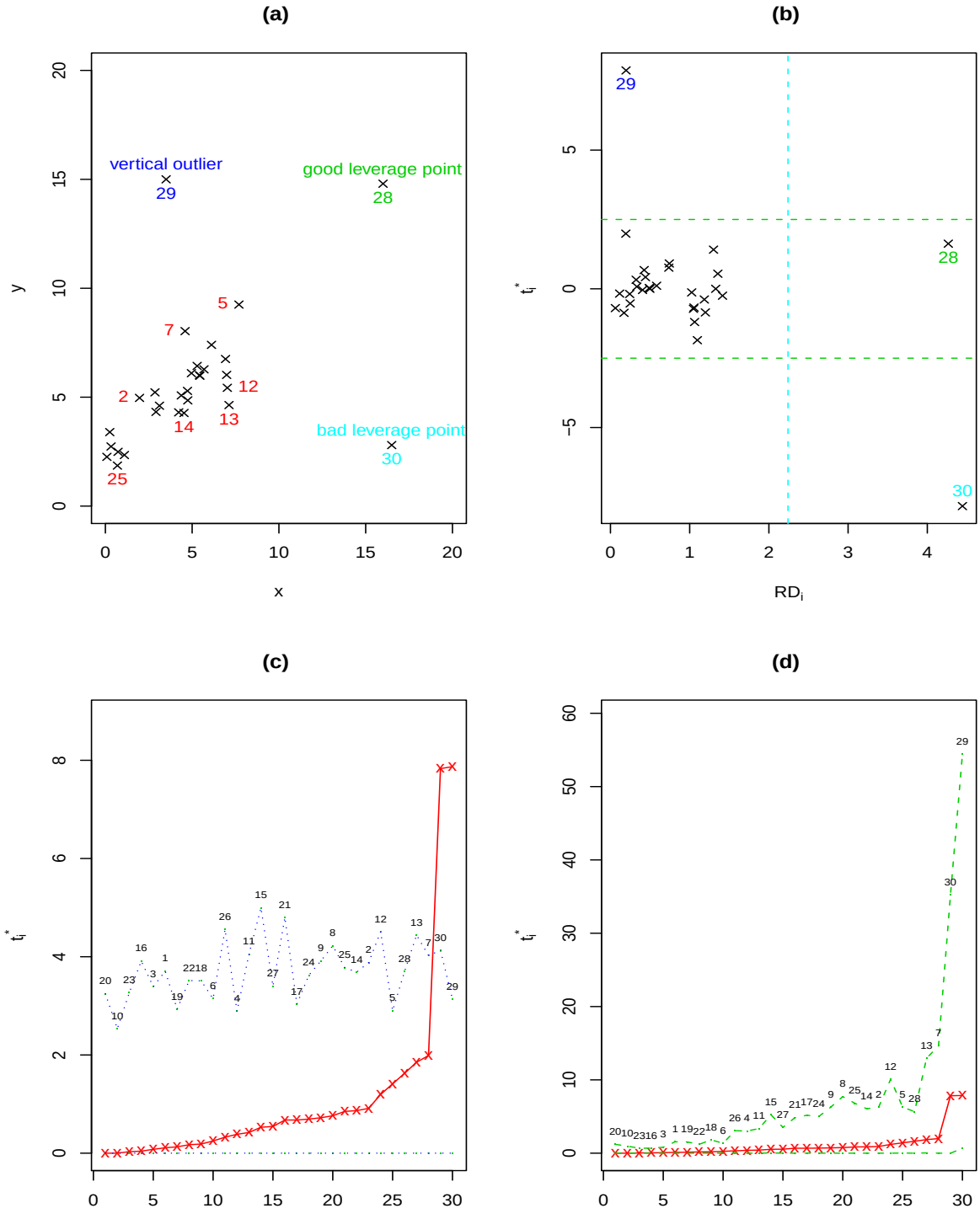


Figure 1: Data types: (a) scatter plot; (b) diagnostic plot; (c) jigsaw plot using (7); (d) jigsaw plot using (8)

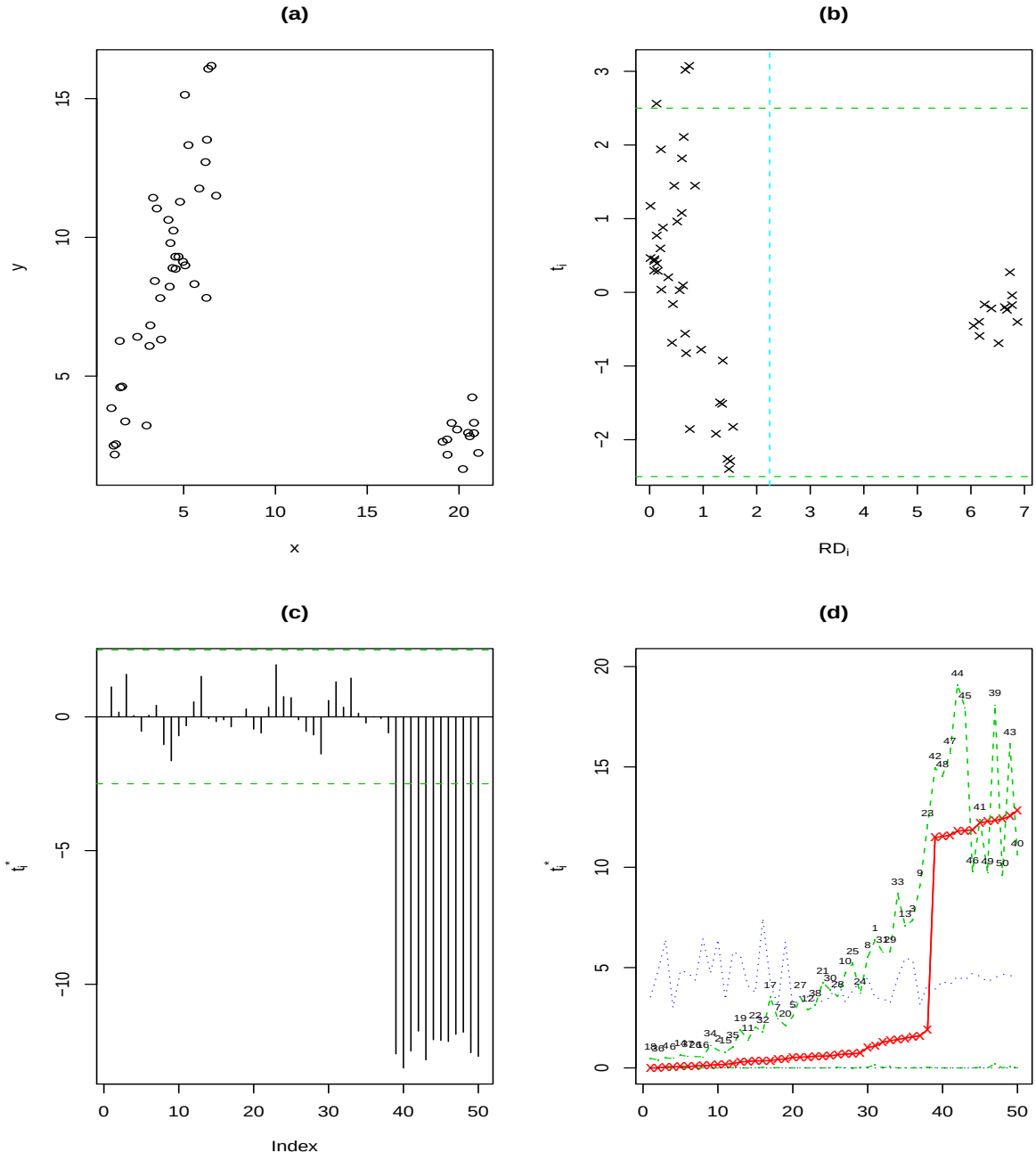


Figure 2: Simulated data with bad leverage points: (a) scatter plot; (b) diagnostic plot based on MM estimate; (b) standardized WLAD residual plot; (d) jigsaw plot.

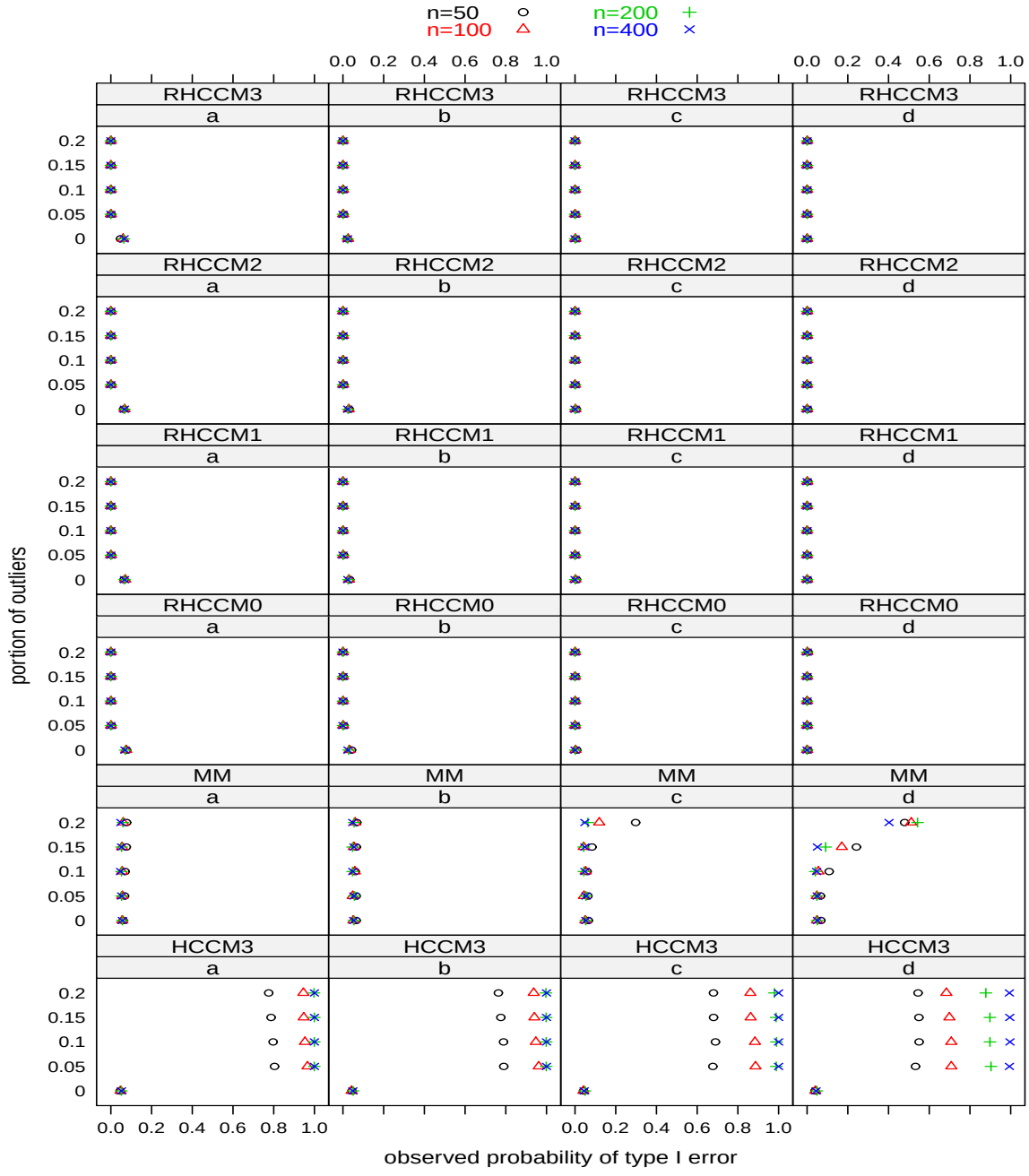


Figure 3: Average of empirical p-values for t tests based on different HCCM estimates and approaches

Monitoring Profile Based on a Linear Regression Model with Correlated Errors

Tsung-Chi Cheng

Department of Statistics
National Chengchi University
64 Zhih-Nan Road, Section 2
Taipei 11623, Taiwan
E-mail: chengt@nccu.edu.tw

Joint work with Su-Fen Yang

Profile monitoring

- Profile monitoring is the use of control charts for cases in which the quality of a process or product can be characterized by a functional relationship between a response variable and one or more explanatory variables.
  - Linear regression model: Kang and Albin (2000); Kim *et al.* (2003); Mahmoud and Woodall (2004); Wang and Tsung (2005); Gupta *et al.* (2006); Mahmoud *et al.* (2006)
  - Nonparametric regression model: Zhou *et al.* (2007)
  - Nonlinear mixed model: Jensen et al. (2008); Jensen and Birch (2009)
  - General: Woodall *et al.* (2004); Woodall (2007)
  - A simple linear profile with AR(1) error: Noorossana *et al.* (2008); Soleimani *et al.* (2009)

Linear regression model with ARMA errors

- Consider the linear regression model

$$y_t = \boldsymbol{x}_t^T \boldsymbol{\beta} + \epsilon_t, \quad t = 1, 2, \cdots, T \tag{1}$$

where y_t is the response variable, $\boldsymbol{x}_t$ is $k \times 1$ vector of explanatory variables, and $\boldsymbol{\beta}$ is a vector of unknown parameters.

- The random error ϵ_t follows an ARMA process, which can be expressed as

$$\Phi(B)\epsilon_t = \Theta(B)\nu_t, \tag{2}$$

where

$$\begin{aligned} \Theta(B) &= 1 + \theta_1 B + \ldots + \theta_q B^q \\ \Phi(B) &= 1 - \varphi_1 B - \varphi_2 B^2 - \ldots - \varphi_p B^p \end{aligned}$$

and $\nu_t \sim WN(0, \sigma^2)$.

Maximum likelihood estimation

- The computation for the estimates can be easily implemented by means of converting models (1) and (2) into the state space form and applying the Kalman filter recursive approach.
- See Harvey (1989); Durbin and Koopman (2001).
- arima function in R.

Hotelling's T^2 test for coefficients

- To monitor the departure of coefficients, $\boldsymbol{\delta} = \{\beta_0, \beta_1, \cdots, \beta_k, \phi_1, \cdots, \phi_p, \theta_1, \cdots, \theta_q\}$, from the profile models (1) and (2) applied to m datasets, the analogous Hotelling's T^2 test is then

$$T_{1i}^2 = (\hat{\boldsymbol{\delta}}_i - \bar{\boldsymbol{\delta}})^T \Delta^{-1} (\hat{\boldsymbol{\delta}}_i - \bar{\boldsymbol{\delta}}), i = 1, 2, \cdots, m,$$

  where $\hat{\boldsymbol{\delta}}_i$ denotes the estimate of $\boldsymbol{\delta}$ for the ith dataset, $\bar{\boldsymbol{\delta}}$ entails the averages of all $\hat{\boldsymbol{\delta}}_i$'s, and $\Delta = \sum_{i=1}^{m} (\hat{\boldsymbol{\delta}}_i - \bar{\boldsymbol{\delta}})(\hat{\boldsymbol{\delta}}_i - \bar{\boldsymbol{\delta}})^T / (m-1)$.

- The $100(1-\alpha)$ percentile of the F distribution can be used to construct an upper control limit (UCL) represented by
  - $\frac{r(T-1)}{(T-r)} F_{\alpha,r,T-r}$ for phase I control chart
  - $\frac{r(T+1)(T-1)}{T(T-r)} F_{\alpha,r,T-r}$ for phase II monitoring scheme
  - $r = k + p + q + 1$

T^2 test based on residuals

- If $\boldsymbol{e}_i$ denotes the $T \times 1$ residual vector for the ith dataset and $\hat{\sigma}_i^2$ is the corresponding estimate of σ^2 for dataset i , then we check the stability of the variance, σ^2, in the profile using the following test statistic,

$$T_{2i}^2 = (\boldsymbol{e}_i - \mathbf{0})^T \Sigma_e^{-1} (\boldsymbol{e}_i - \mathbf{0}), i = 1, 2, \cdots, m,$$

  where $\Sigma_e = \bar{\sigma}^2 I$, $\bar{\sigma}^2$ is the average of all $\hat{\sigma}_i^2$'s, and I is the identity matrix.

- The UCL for this test statistic is $\chi^2_{\alpha, T-1}$, which denotes the $100(1-\alpha)$ percentile of the χ^2 distribution with $T-1$ degrees of freedom.

Simulation study

- Consider the following model

$$\begin{aligned} y_t &= \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t}, t = 1, 2, \cdots, T, \\ \epsilon_t &= \phi_1 \epsilon_{t-1} + \phi_2 \epsilon_{t-2} + \phi_3 \epsilon_{t-3} + \nu_t \end{aligned}$$

  where $\nu_t \sim WN(0, \sigma^2)$.

- Given that $\beta_0 = \beta_2 = 1$, $\phi_2 = 0.1$ and $\phi_3 = -0.1$, we focus on evaluating the impact of changes in β_1, ϕ_1, and σ^2 on the monitoring profile.
- There are 50000 replicates used for Phase II diagnostic monitoring, while 20000 replicates are carried out for Phase I control chart scheme.
- For the latter, both values of β_1 and σ^2 are assigned to be 1, while the values of ϕ_1 vary from -0.6 and 0.6 to avoid non-stationary series occurring in the data generating process.
- The sample size, T, is 150 and 300.

In-control ARL

- The combination of T_1^2 and T_2^2 control chart schemes is considered to yield an overall in-control average run length (ARL) of approximately 185.
- The overall in-control ARL can be calculated by $1/ARL_{overall} = 1 - (1 - \alpha_1)(1 - \alpha_2)$, where α_1 and α_2 denote the probability of committing false alarms for T_1^2 and T_2^2, respectively.

The simulated ARL values under the change of β_1 from 1

	ϕ_1	β_1					
		1	1.02	1.04	1.06	1.08	1.10
$T = 150$	-0.6	**145.349**	19.231	1.588	1.007	1.000	1.000
	-0.4	**158.228**	29.851	2.249	1.044	1.000	1.000
	-0.2	**178.571**	46.685	3.618	1.194	1.004	1.000
	0.0	**175.439**	66.225	7.198	1.663	1.053	1.001
	0.2	**171.233**	79.872	13.221	2.685	1.257	1.022
	0.4	**190.114**	107.527	25.253	5.247	1.882	1.164
	0.6	**170.068**	125.313	40.783	10.156	3.181	1.573
$T = 300$	-0.6	**177.305**	4.224	1.011	1.000	1.000	1.000
	-0.4	**188.679**	6.973	1.068	1.000	1.000	1.000
	-0.2	**177.936**	11.743	1.249	1.001	1.000	1.000
	0.0	**188.679**	20.309	1.750	1.016	1.000	1.000
	0.2	**187.266**	34.916	2.962	1.116	1.001	1.000
	0.4	**194.553**	54.171	5.503	1.468	1.031	1.000
	0.6	**181.159**	68.213	9.750	2.190	1.152	1.010

The simulated ARL values under the varying values of ϕ_1

T	ϕ_1	ϕ_1						
		0	-0.1	-0.2	-0.3	-0.4	-0.5	-0.6
150	-0.6	1.000	1.001	1.024	1.283	3.089	19.216	**147.929**
	-0.4	1.149	1.968	6.539	43.365	**180.505**	112.867	16.706
	-0.2	10.301	60.976	**159.236**	108.225	20.938	3.544	1.305
300	-0.6	1.000	1.000	1.000	1.003	1.282	7.504	**142.450**
	-0.4	1.000	1.056	2.058	17.094	**177.936**	39.777	2.302
	-0.2	2.802	27.337	**189.394**	44.326	3.294	1.111	1.001
		0	0.1	0.2	0.3	0.4	0.5	0.6
150	0.0	**175.439**	78.125	12.994	2.935	1.317	1.032	1.001
	0.2	13.203	76.453	**149.254**	77.760	12.713	2.621	1.206
	0.4	1.260	2.460	9.651	60.168	**166.113**	79.365	10.231
	0.6	1.000	1.006	1.094	1.709	5.572	38.580	**163.399**
300	0.0	**188.679**	33.201	3.043	1.142	1.002	1.000	1.000
	0.2	3.128	32.489	**171.821**	34.626	2.903	1.097	1.001
	0.4	1.001	1.094	2.511	24.606	**191.571**	31.726	2.169

The simulated ARL values under the shifts of σ from 1

		σ					
	ϕ_1	1	1.1	1.2	1.3	1.4	1.5
$T = 150$	-0.6	**145.349**	5.677	1.415	1.034	1.001	1.000
	-0.4	**158.228**	5.578	1.399	1.034	1.001	1.000
	-0.2	**178.571**	5.648	1.410	1.033	1.002	1.000
	0.0	**175.439**	5.752	1.411	1.033	1.002	1.000
	0.2	**171.233**	5.633	1.407	1.032	1.001	1.000
	0.4	**190.114**	5.476	1.399	1.032	1.001	1.000
	0.6	**170.068**	5.725	1.412	1.033	1.001	1.000
$T = 300$	-0.6	**177.305**	2.619	1.040	1.000	1.000	1.000
	-0.4	**188.679**	2.571	1.040	1.000	1.000	1.000
	-0.2	**177.936**	2.588	1.040	1.000	1.000	1.000
	0.0	**188.679**	2.614	1.041	1.000	1.000	1.000
	0.2	**187.266**	2.605	1.040	1.000	1.000	1.000
	0.4	**194.553**	2.582	1.041	1.000	1.000	1.000
	0.6	**181.159**	2.618	1.041	1.000	1.000	1.000

Conclusions from simulation

- The test statistics are sensible to the alterations of the coefficients, namely β_1, ϕ_1, and σ^2.
- The ARL value is more sensible for the difference when $T = 150$ than that for $T = 300$.
- The detection of the change in β_1 may depend on the values of ϕ_1.
  - The pattern about how this differs may depend on more factors, such as the complexity of error function (2) and/or shifts in different parameters simultaneously.
  - To verify this, more simulations should be expected.

Real data illustration

- The “Babyfinder” is a device designed to detect if any event of concern occurs.
  - For example, it could be widely used in healthcare, warehouse, for baby sisters, heart trouble patients, stolen bicycles, etc.
- It includes transceiver and receiver.
  - Once there is a distance between transceiver and receiver, a signal strength is generated.
  - The signal strength is called Received Signal Strength Indicator (RSSI), a measurement of the power presents in a received radio signal (measured in decibels, dBs), in wireless communication technology.
  - In wireless communication theory, the functional relationship between RSSI and distance should be expressed by the model:

    $$\text{RSSI} = a + b \log(\text{distance}),$$

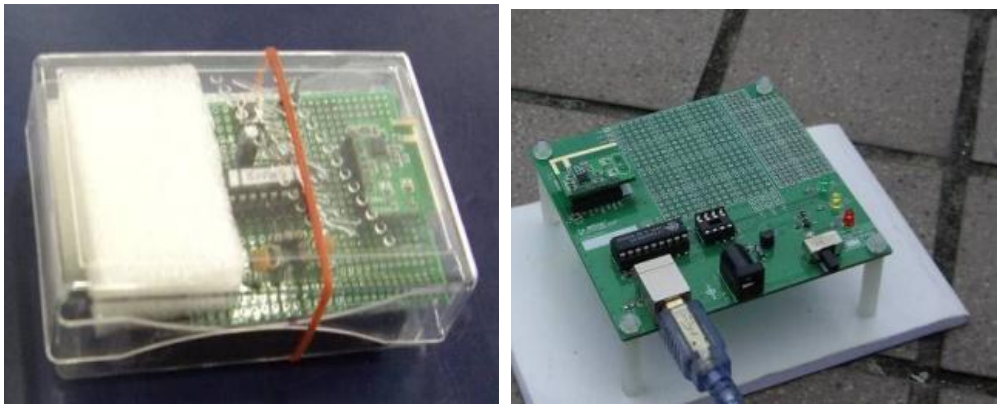
    where a is the intercept and b is the slope.

Phase I control chart

- The study problem of Babyfinder is to analyze the behavior of the wireless signal strength through various action models.
  - The occurring events would change the functional relationship of RSSI and distance.
  - Hence, it is important to effectively detect if the functional relationship of RSSI and distance has changed.
- Suppose that the Babyfinder is developed to protect a bicycle from being stolen.
  - To collect the data of RSSI under specified distances, seventeen no-stolen experiments (or in-control experiments) are designed, which result from different situations and environments.
- The following regression model is first applied to analyze these datasets:

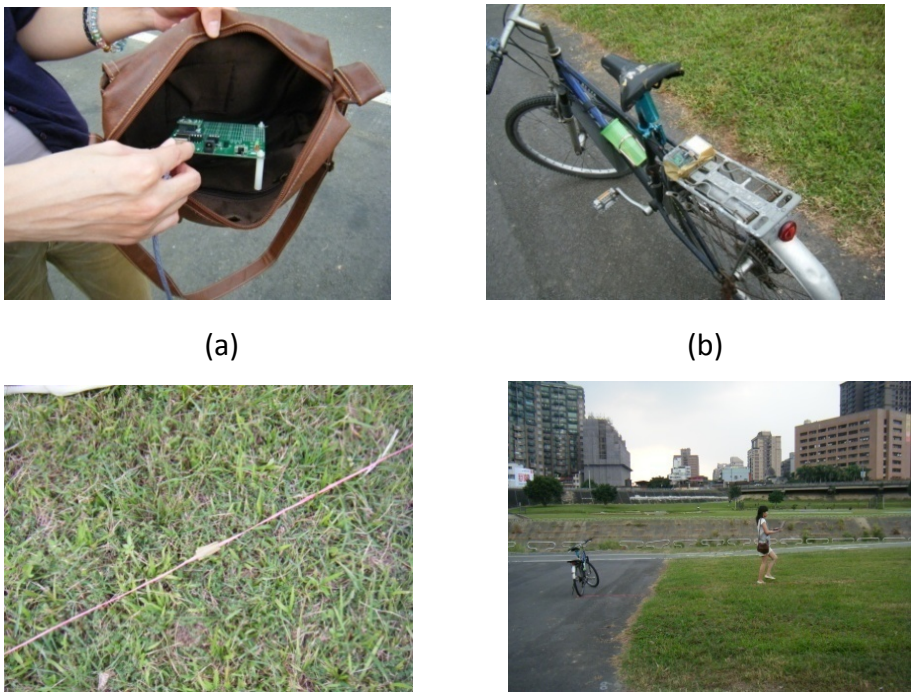
$$y_t = \beta_0 + \beta_1 \log(x_t) + \epsilon_t, \quad t = 1, 2, \cdots, 147,$$

where y_t is RSSI and x_t is the distance between the bicycle and the owner.



(a) (b)

Figure 1. Babyfinder: (a) Transceiver and (b) receiver

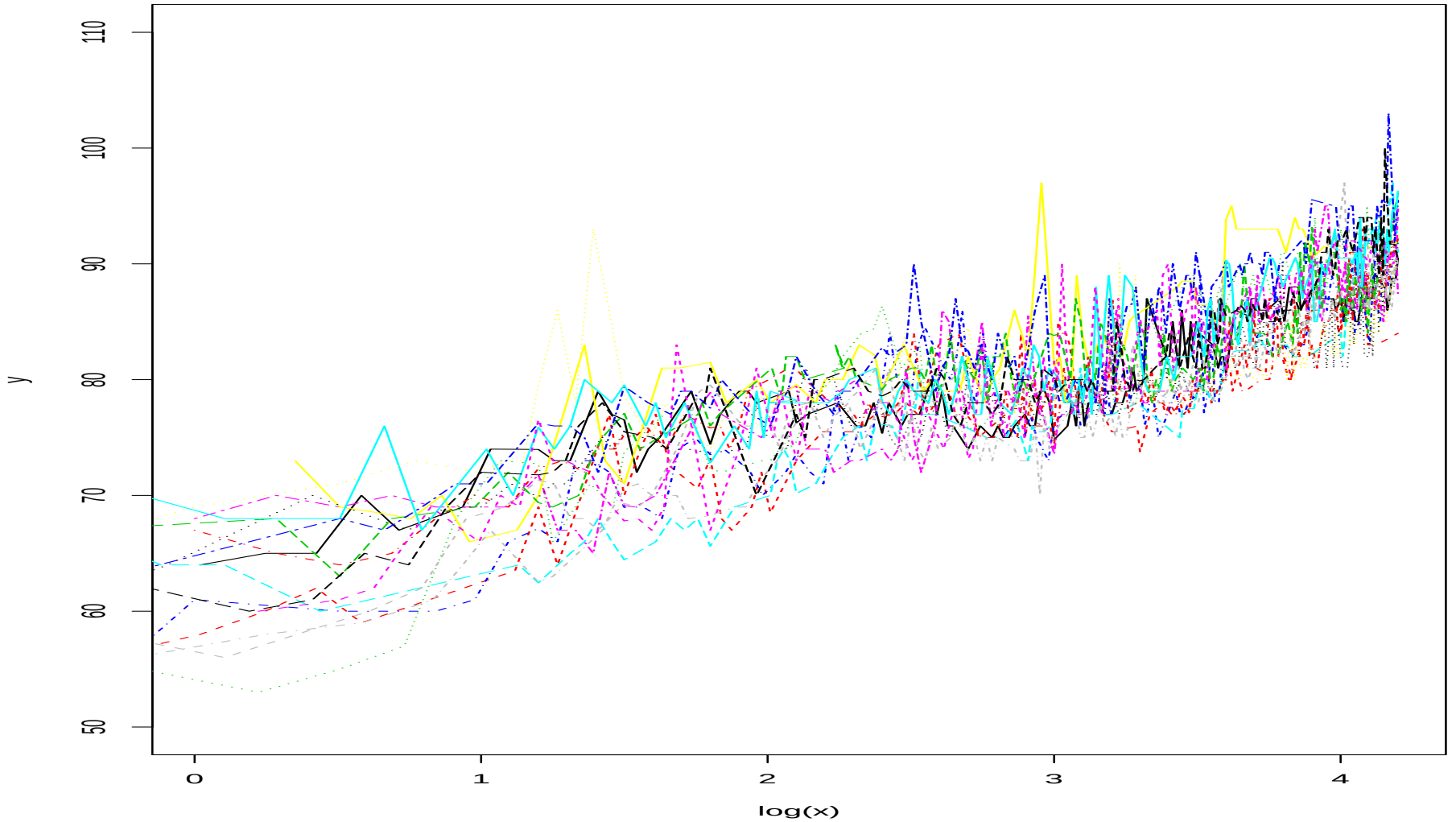


(a) (b)

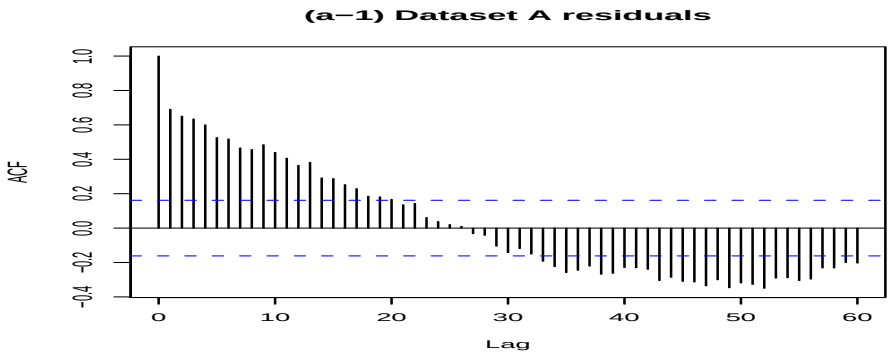
(c) (d)

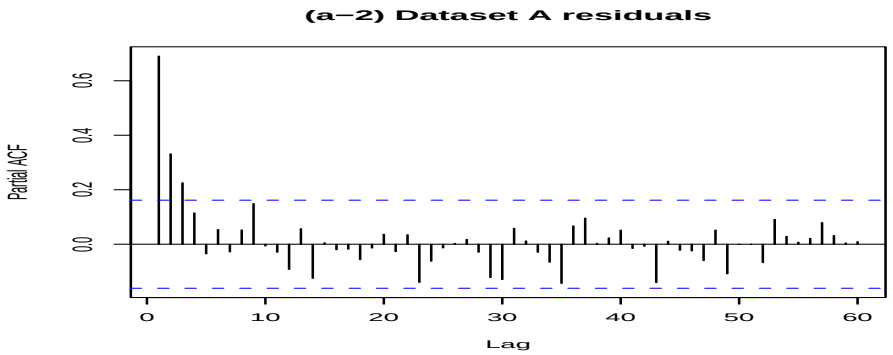
Figure 2. The pictures of an experiment: (a) the receiver is in the owner's bag; (b) the transceiver in on the bicycle; (c) the rope and the marked distances; (d) the owner leaves the bicycle and walks ahead.

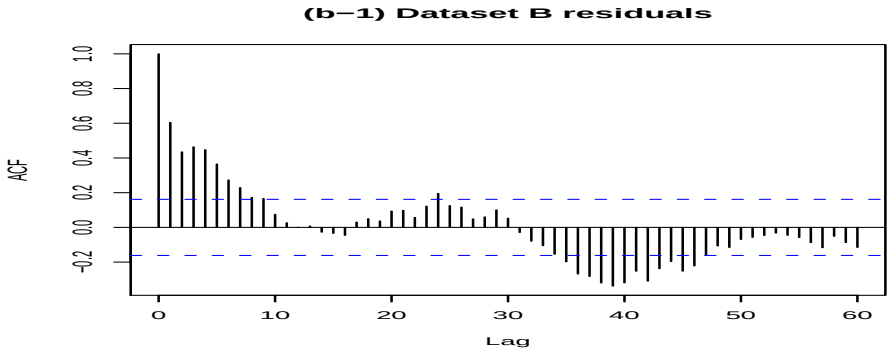
Line charts

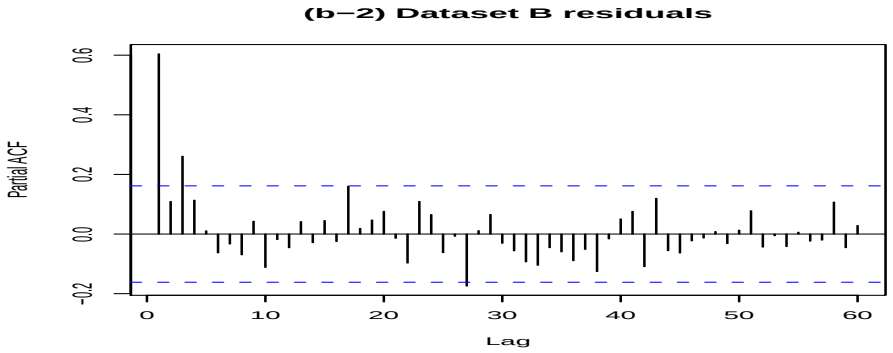


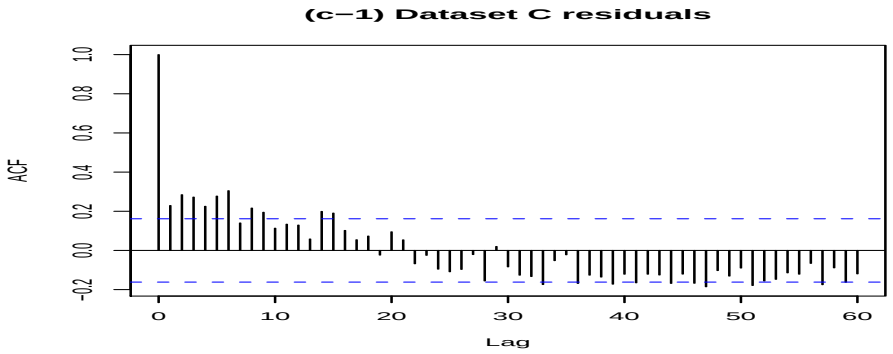
Residuals of fitting a regression with constant errors

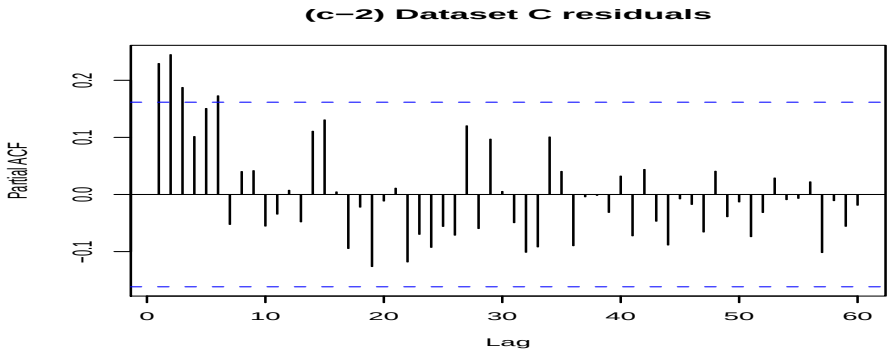












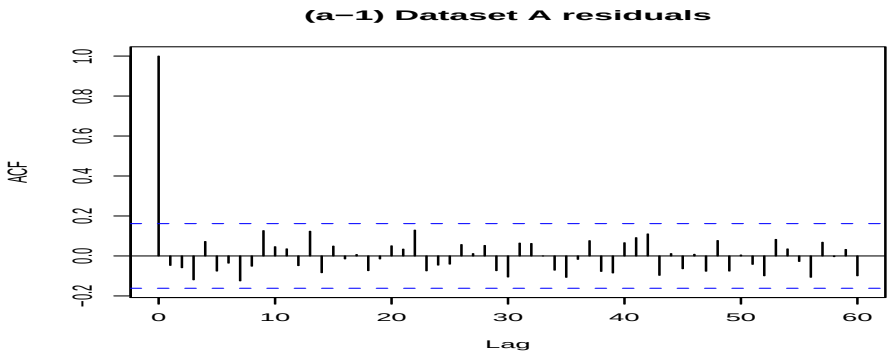
Profile model

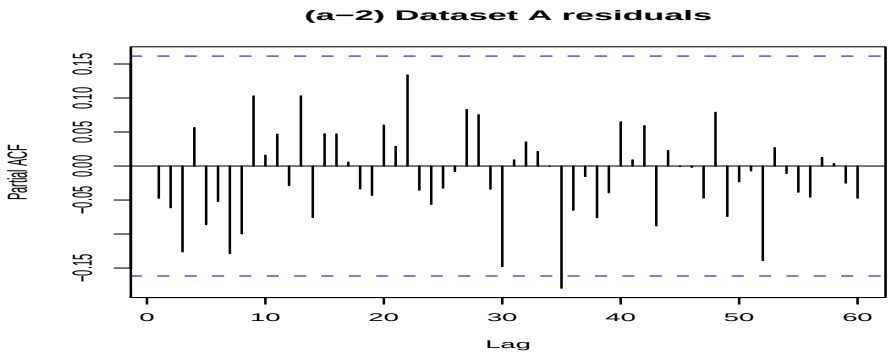
- A candidate profile model is then developed to analyze these datasets:

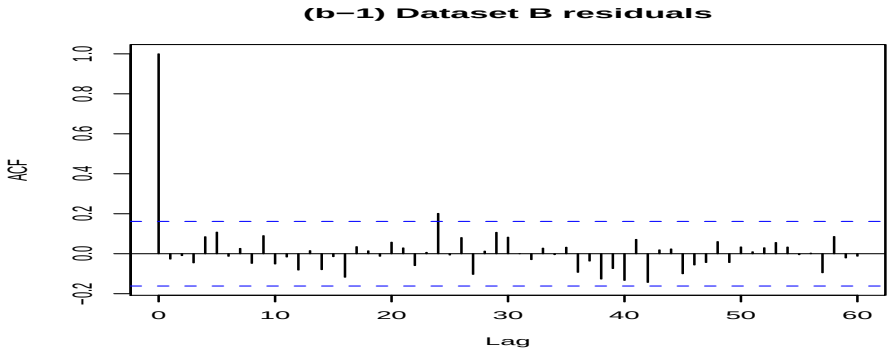
$$
\begin{aligned}
y_t &= \beta_0 + \beta_1 \log(x_t) + \epsilon_t, \quad t = 1, 2, \cdots, 147, \\
\epsilon_t &= \phi_1 \epsilon_{t-1} + \phi_2 \epsilon_{t-2} + \phi_3 \epsilon_{t-3} + \nu_t
\end{aligned}
$$

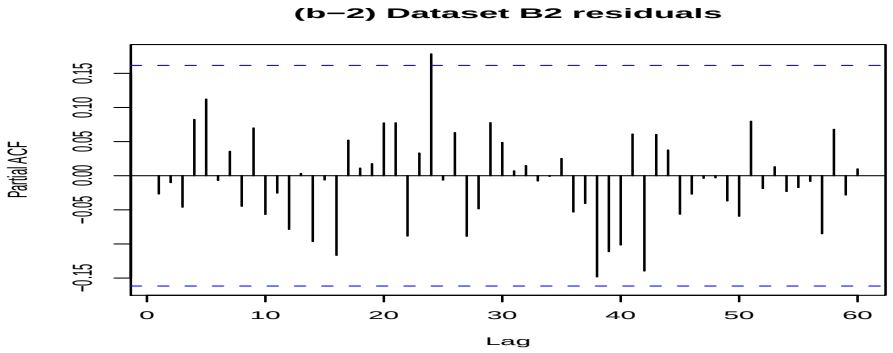
where $\nu_t \sim WN(0, \sigma^2)$.

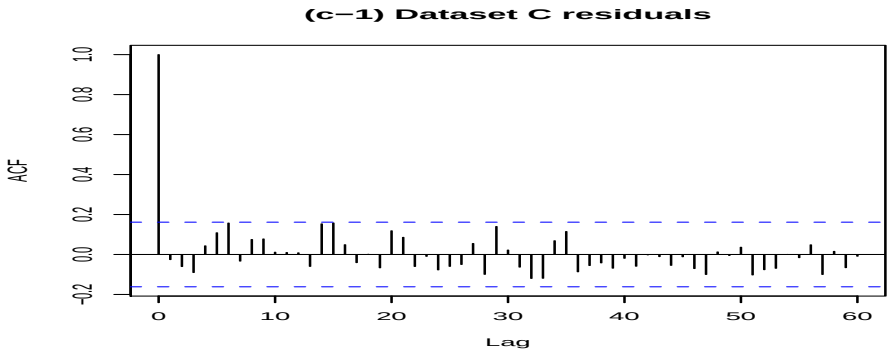
Residuals of fitting a regression with AR(3) errors

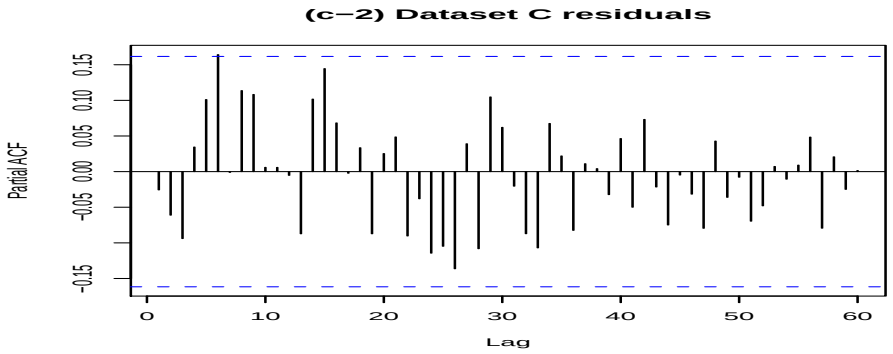




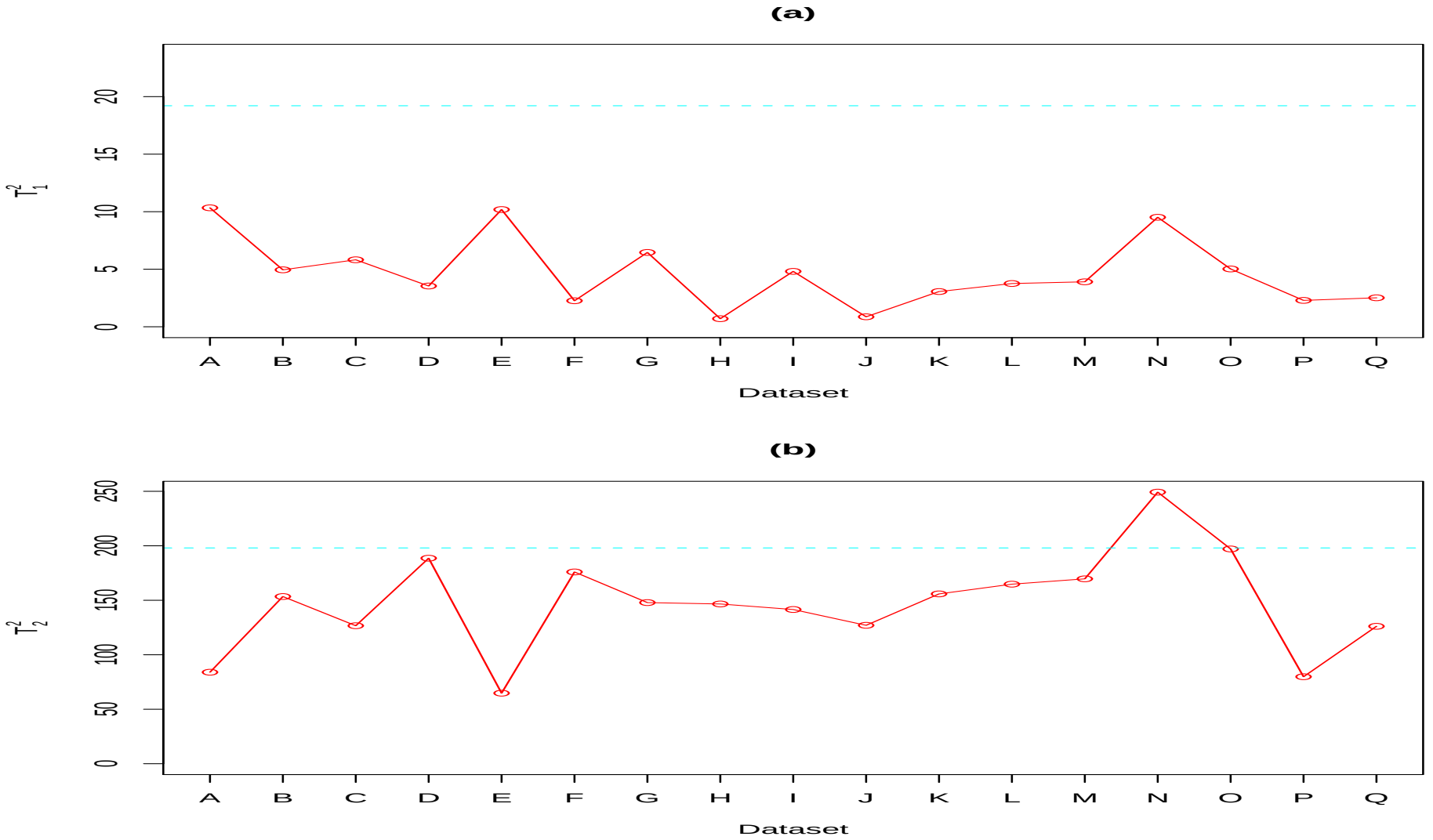




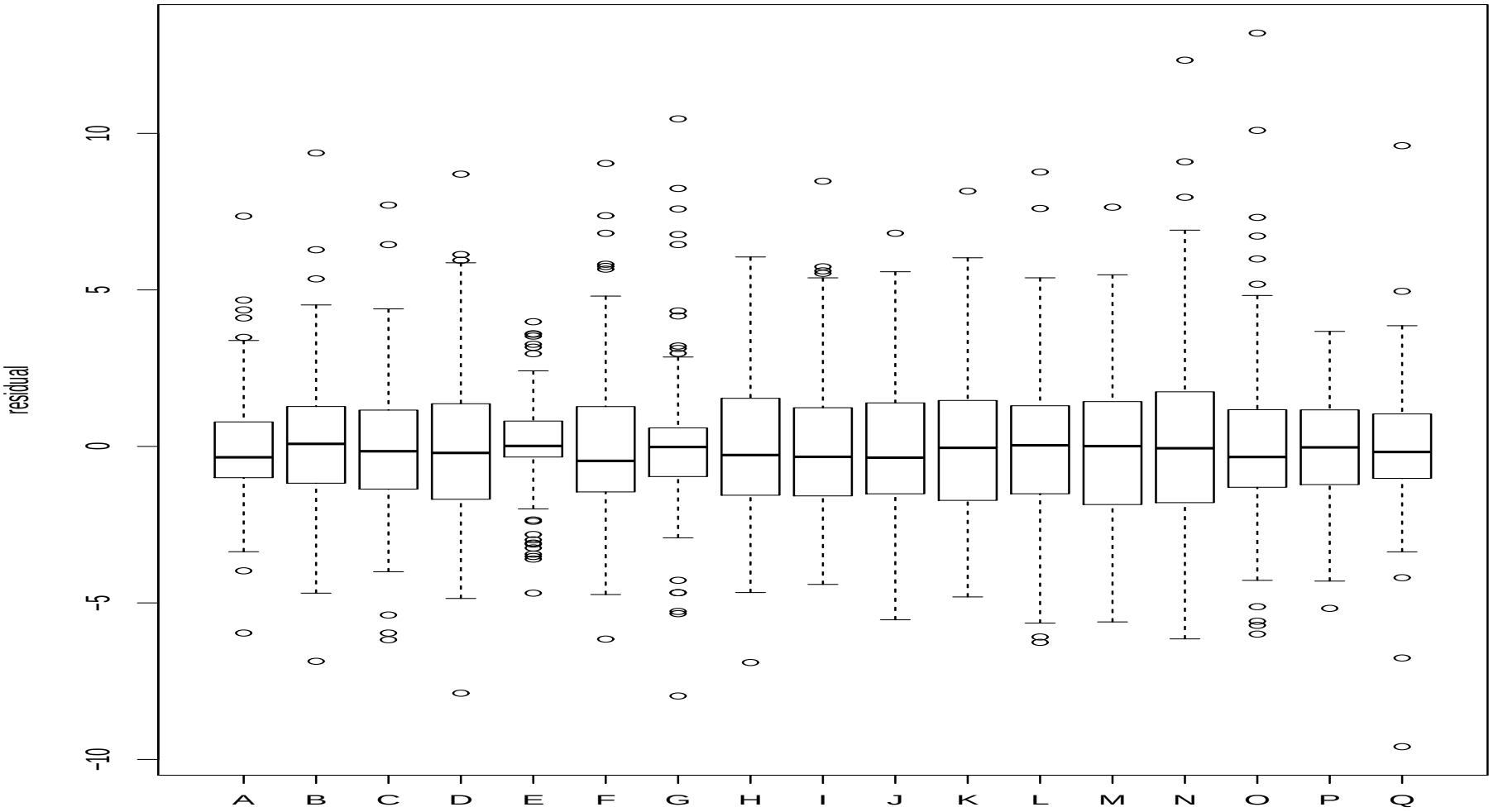




Plots of T^2 statistics



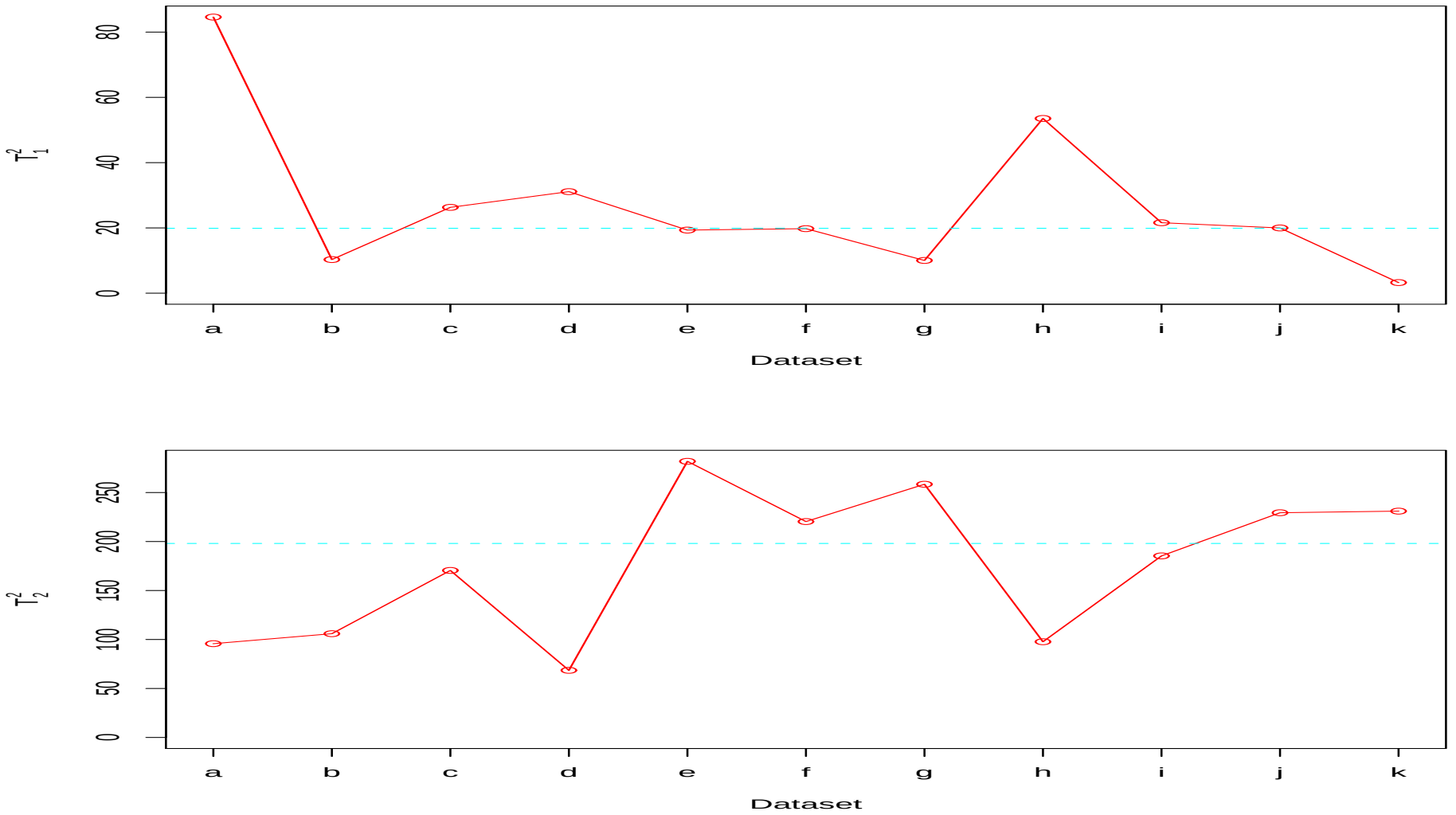
Boxplots of residuals



Phase II monitoring scheme

- To evaluate the capability of the proposed approaches, Eleven stolen experiments (or out-of-control experiments) are designed by the occurring special causes, such as moving speed and methods of stealing.

Plots of T^2 statistics under out-of-control cases



無研發成果推廣資料

98 年度專題研究計畫研究成果彙整表

計畫主持人：鄭宗記	計畫編號：98-2118-M-004-006-
計畫名稱：異質變異數矩陣之穩健估計	

成果項目			量化			單位	備註（質化說明：如數個計畫共同成果、成果列為該期刊之封面故事...等）
			實際已達成數（被接受或已發表）	預期總達成數(含實際已達成數)	本計畫實際貢獻百分比		
國內	論文著作	期刊論文	0	1	100%	篇	
		研究報告/技術報告	0	0	100%		
		研討會論文	1	0	100%		
		專書	0	0	100%		
	專利	申請中件數	0	0	100%	件	
		已獲得件數	0	0	100%		
	技術移轉	件數	0	0	100%	件	
		權利金	0	0	100%	千元	
	參與計畫人力（本國籍）	碩士生	0	0	100%	人次	
		博士生	0	0	100%		
		博士後研究員	0	0	100%		
		專任助理	0	0	100%		
國外	論文著作	期刊論文	0	0	100%	篇	
		研究報告/技術報告	0	0	100%		
		研討會論文	0	0	100%		
		專書	0	0	100%	章/本	
	專利	申請中件數	0	0	100%	件	
		已獲得件數	0	0	100%		
	技術移轉	件數	0	0	100%	件	
		權利金	0	0	100%	千元	
	參與計畫人力（外國籍）	碩士生	0	0	100%	人次	
		博士生	0	0	100%		
		博士後研究員	0	0	100%		
		專任助理	0	0	100%		

其他成果 (無法以量化表達之成果如辦理學術活動、獲得獎項、重要國際合作、研究成果國際影響力及其他協助產業技術發展之具體效益事項等，請以文字敘述填列。)	無

	成果項目	量化	名稱或內容性質簡述
科教處計畫加填項目	測驗工具(含質性與量性)	0	
	課程/模組	0	
	電腦及網路系統或工具	0	
	教材	0	
	舉辦之活動/競賽	0	
	研討會/工作坊	0	
	電子報、網站	0	
	計畫成果推廣之參與（閱聽）人數	0	

國科會補助專題研究計畫成果報告自評表

請就研究內容與原計畫相符程度、達成預期目標情況、研究成果之學術或應用價值（簡要敘述成果所代表之意義、價值、影響或進一步發展之可能性）、是否適合在學術期刊發表或申請專利、主要發現或其他有關價值等，作一綜合評估。

1. 請就研究內容與原計畫相符程度、達成預期目標情況作一綜合評估

■達成目標

□未達成目標（請說明，以 100 字為限）

□實驗失敗

□因故實驗中斷

□其他原因

說明：

2. 研究成果在學術期刊發表或申請專利等情形：

論文：□已發表 ■未發表之文稿 □撰寫中 □無

專利：□已獲得 □申請中 ■無

技轉：□已技轉 □洽談中 ■無

其他：（以 100 字為限）

已將本研究成果撰寫成論文送至國際期刊審稿中。

3. 請依學術成就、技術創新、社會影響等方面，評估研究成果之學術或應用價值（簡要敘述成果所代表之意義、價值、影響或進一步發展之可能性）（以 500 字為限）

本研究藉由加權最小絕對離差(weighted least absolute deviation，WLAD)估計法，成功地利用圖形方式區分資料之異質性與離群值；就文獻及方法而言，有其極大的突破與貢獻。本文之結果，在資料分析上提供一個有利的工具；預期本文將可於統計計算方面的國際期刊發表。