AUTOMATION OF LIKELY OUTLIERS DETECTION IN LINEAR MIXED MODELS

by

Yue Wang

A thesis submitted to the Faculty of the University of Delaware in partial fulfillment of the requirements for the degree of Master of Science in Statistics

Fall 2013

AUTOMATION OF LIKELY OUTLIERS DETECTION

IN LINEAR MIXED MODELS

by

Yue Wang

Approved: _____
Jong Soo Lee, Ph.D.
Co-Professor in charge of thesis on behalf of the Advisory Committee

Approved: _____
Randall J. Wisser, Ph.D.
Co-Professor in charge of thesis on behalf of the Advisory Committee

Approved: _____
Titus Awokuse, Ph.D.
Chair of the Department of Food and Resource Econmics

Approved: _____
Mark Rieger, Ph.D.
Dean of the College of Agriculture and Natural Resources

Approved: _____
James G. Richards, Ph.D.
Vice Provost for Graduate and Professional Education

ACKNOWLEDGMENTS

Firstly, I would like to express my deep appreciation to my advisor, Dr. Jong Soo Lee. Without his guidance and persistent help, this thesis would not have been possible. Dr. Lee's encouragement, advice and patience have motivated me to make continuous progress in this research project. Also, I really enjoy Dr. Lee's class STAT601, which is so well organized, informative, interesting, challenging but not overwhelming. I feel honored to have him as my advisor.

I also would like to express my sincere gratitude to my co-advisor Dr. Randall J. Wisser. I am inspired by his intelligence, diligence, efficiency and spirit of adventures and curiosities in research. I learnt a lot from him each time in our meeting. Besides spending his valuable time to give me guidance, Dr. Wisser also helped in running my code on his server, even on weekends and holidays. I am thankful to his great help.

I would like to thank Dr. John D. Pesek for being my committee member. Actually I regard him partially as committee member and partially as an advisor because of the time he spent to have meetings with me and teach me about the knowledge that I need in completing this study. I gained most of my knowledge about linear mixed model from Dr. Pesek. Also, his classes STAT 674, 675 and 616 provide students with solid foundations they need in their future careers, which his students would appreciate after their graduations.

I also would like to thank Dr. Thomas W. Ilvento who admitted me to Statistics program. I am thankful for his hard work in helping students in polishing

resumes, finding internships, preparing students for interviews and encouraging students when they are frustrated.

I thank Zaiqi Pan, Bruce H Stanley and David W. Onstad in DuPont Pioneer, where I interned, for their support in my research project and their flexibility when I need to ask for leave to have a meeting with my advisors.

Thanks to all my friends, especially Miao Tang, Lijun Chen, Duo Huang, Shuyan Li, Jing Li and Lei Zheng for the countless happy hours we had together, on campus or off-campus. You made my experience of the two and a half years' study in Statistics program more enjoyable. I have a wonderful memory of this learning and growing experience because of you all.

Last but not the least; let me express my deepest gratitude and sincerest appreciation to my parents, my husband and my other family members for always loving me, caring for me and supporting me!

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Key words: Outliers Detection, Linear Mixed Models, Simulation and Automation.

It is difficult to detect outliers in linear mixed models. The traditional way of identifying outliers is to check whether there are any violations in model assumptions by examining the normal QQ plot and the residual plot. A simulation approach proposed by Schützenmeister and Piepho adds the objectivity in interpreting results of the QQ and residual plot. Based on this simulation approach, a software tool is developed to indentify potential outliers in linear mixed models automatically. In addition, the performance of this approach is evaluated. This tool is user-friendly to inexperienced analysts and open sourced.

Chapter 1

INTRODUCTION

The statistical methodologies have improved significantly over the last several decades. The advances of those methodologies have enabled agricultural and natural resources science developed dramatically. Among those methodologies, the generalized linear mixed model has made exceptional contributions in facilitating researchers to conduct more versatile and informative analyses. The development of user-friendly statistical software benefits a wider range of researchers in allowing them to utilize generalized linear mixed models in their analysis, the access of which was limited in the past (Gbur, et al., 2012).

When analyzing an agriculture data set with a generalized linear mixed model, an important step is the outlier diagnostics. The purpose of this study is to develop an open-sourced software tool, which identifies the potential outliers in an automatic manner. The tool is developed with R, and is accessible to even smallholder farmers in developing countries. It helps them to obtain useful information with their limited budget. The tool is automatic and user friendly for even inexperienced users.

A linear mixed model is written as (Littell et al., 2006):

$$Y = X\beta + Zu + e,$$

$$u \sim N(0, G),$$

$$e \sim N(0, R),$$

$$Cov[u, e] = 0,$$

where β is the fixed effects vector and u the random effects. X is the design matrix for the fixed effects, Y is the vector of responses and Z is the design matrix for the random effects. The unknown parameters of the model to be estimated are β, G and R.

A linear mixed model is an extension of a general linear model, which is in the form of:

$$Y = Xb + e$$

$$e \sim N(0, \sigma^2)$$

Not only has the error term e as a random component as that in a general linear model, linear mixed models also incorporate the random effect u, which brings in many advantageous properties. It is capable of making a broader inference to different environments and modeling non-independent datasets. As a result, the general linear mixed model is widely applicable in various fields, such as plant breeding or a longitudinal data analysis (Schutzenmeister and Piepho, 2012).

In general linear models, the residual is defined as: $\hat{e} = Y - X\hat{b}$, which is the difference between the observed response variable and the estimated response variable. In contrast, there are three types of residuals in linear mixed models (Nobre and Singer, 2007):

(1) the marginal residuals,

$\hat{\xi} = y - X\hat{\beta}$, which is the difference between the observed distribution y and the estimated marginal distribution $X\hat{\beta}$. The marginal distribution of y is obtained by "integrating the joint distribution of data and random effects over random effects" (Littell, 2006) and has a mean $X\beta$ and variance $ZGZ'+R$.

(2) the conditional residuals,

$\hat{\varepsilon} = y - X\hat{\beta} - Z\hat{u}$, which is the difference between the observed distribution and the estimated conditional distribution $X\beta + Zu$. The conditional distribution of $Y \mid u$ is obtained by giving Zu as a known fixed constant. $Y \mid u$ has a mean of $X\beta + Zu$ and variance R.

(3) best linear unbiased predictor (BLUP), $zu = E[y \mid u] - E[y]$, which predicts the random effects.

Different types of residuals serve for different diagnostic purposes. For instance, the marginal residuals could diagnose the linearity of effects as well as the within subjects covariance matrix, the conditional residual could test for outlying observations and homoscedasticity and normality of conditional errors, whereas the BLUP can test the presence of outlying subjects, random effects covariance structure and normality of the random effects. Residuals, especially studentized residuals are essentially important in detecting outliers. A data point with a greater (studentized) residual is more likely to be considered as an outlier.

Outliers, according to Gumedze et al (2010), are "data observations that fall outside the normal range of the response data". Whether or not an observation is an outlier may be dependent on the selected variables in the model (Nobre and Singer 2007). Thus, outlier detection unavoidably involves some uncertainties due to the indetermination in the parameter estimation and model selection. Therefore, it is inappropriate to connect some certainty-implying terms such as identification and detection with outliers. In fact, it's more appropriate to refer "outlier identification" as "likely outlier identification" (Longford, 2001).

There is a formal test for outliers in linear models (Kutner et al., 2004), which involves residual studentization. Studentization of residual means the division of a

residual by an estimate of its standard deviation. The reason for conducting studentization is that the standard deviations of residuals in a sample often vary significantly for different data points. Thus the residuals are not comparable without studentization. The studentized residual, with the i[th] element deleted, is calculated as:

$$t_i = \frac{d_i}{s\{d_i\}} = \frac{e_i}{\sqrt{MSE_{(i)}(1-h_{ii})}} \sim t(n-P-1)$$

where h_{ii} is the ith diagonal element in the hat matrix H. The hat matrix is defined as: $H = X(X^TX)^{-1}X^T$. It is only dependent upon the design matrix X and can be regarded as the orthogonal projection onto the column space of X. Since t_i follows the t distribution with n-p-1 degrees of freedom, the t test is therefore able to test if an observation is an outlier or not.

After outlying points are detected, sometimes we need to estimate the influence of those outlying points. One of the popular measures is Cook's Distance, which evaluates the influence of the ith case on all the fitted values ((Kutner et al., 2004). It is defined as:

$$D_i = \frac{\left(\hat{\mathrm{Y}} - \hat{\mathrm{Y}}_{(i)}\right)'\left(\hat{\mathrm{Y}} - \hat{\mathrm{Y}}_{(i)}\right)}{pMSE},$$

where p is the number of regression parameters, $\hat{\mathrm{Y}}$ is the vector of the fitted values when all the cases are included and $\hat{\mathrm{Y}}_{(i)}$ is the vector of the fitted values when the ith case is omitted in the regression. When relating D_i to the F(p,n-p) distribution, if the corresponding percentile value is near or greater than 50% , it indicates the ith case has a big influence.

Unfortunately, this outlier diagnostic analysis cannot be simply extended to linear mixed models, and there actually is no valid test for identification of outliers in

linear mixed models (Andrade-Bejarano and Longford, 2010). Some outliers may be resulted from incorrect data transcription or errors of experimental equipments and can be easily detected and removed (Freedom et al., 2010). However, more often the outliers have undetectable sources and are difficult to deal with, since they cannot be simply discarded without convincing reasons. The remedy approaches to outliers with unknown origin include data transformation or model modification. Despite of those remedies, we still never know whether the detected outliers are true outliers or the assumptions of our model are incorrect in the first place.

A common way to test for outliers in linear mixed models is to generate conditional and marginal residual plots and see if the residuals appear to be normal and homoscedastic. However, it cannot avoid some level of subjectivity (Schutzenmeister and Piepho, 2012). Two other main approaches are case-deletion and influence analysis (Gumedze et al. 2010). The disadvantage of the case-deletion method is that the observations considered outliers ended up being discarded even when they are not true outliers, as a result, some useful information may have been lost. On the other hand, influence analysis is to test "how influential a designed point is to a certain model". An unreasonable influential observation should be avoided by a good experimental design and it is out of the scope of our work.

Due to the difficulty of detecting outliers in linear mixed models, the purpose of this work is not to propose a ground-breaking method to determine whether or not the potential outliers should be discarded. Instead, an approach that automatically detects the likely outliers in mixed models will be illustrated. The approach adopted in the work is proposed by Schützenmeister (2012). It is a simulation-based approach that allows automatic detection of potential outliers in the model. The basic idea is to

assess normality and homoscedasticity of residuals by adding simultaneous tolerance bands (STB) and simultaneously tolerance intervals (STI) on the normal QQ plot and the residual plot. It enables an inexperienced analyst to make judgments with objectivity and does not require deletion of any observations or classify any observations as outliers conclusively. The work done in this thesis includes: programming the R codes to realize the deletion algorithm of the paper by Schützenmeister (2012), conducting simulations to confirm the correctness of the method by comparing the R result with the SAS result, evaluating the performance of this approach and extending the application of the approach to a more complex model.

Chapter 2

SIMULATION APPROACH DESCRIPTIONS

The simulation approach proposed by Schützenmeister and Piepho is to graphically check the normality and homoscedasticity assumptions. One of their assumptions is that there are only a few outliers, which are standing out from the remaining residual points. They made use of the normal QQ plot to test the departure from normality and the conditional studentized residual plot to identify the outstanding data points. By adding the 100(1-α)% STB and STI, they are able to interpret the result objectively. Any data point outside both the STB and STI is regarded as a potential outlier.

One example data set they used in their paper is the Cambridge filter data set. The response variable of the data set is detected nicotine content by gas chromatography. One of the two explanatory variables is the original nicotine content of the Cambridge filter pads sample, which has 10 levels (10 samples). The other is the effect by the labs, where the analyses were conducted. The latter has 14 levels (14 labs).

The linear mixed model to fit the data set is:

$$y_{ij} = \mu + a_i + b_j + e_{ij}, \qquad (1)$$

where y_{ij} (i=1,…,10;j=1,…,14) is the amount of detected nicotine content in the ij-th sample, a_i is the fixed effect of the i-the sample, b_j is the random effect of the j-th lab, and e_{ij} is the error term for the ij-th measurement.

$b_j \sim N(0,\sigma_b^2)$
$e_{ij} \sim N(0,\sigma_e^2)$

There are two missing values in the data set: the 8-th and 9-th sample in the F lab. Thus, there were only 138 measurements.

To describe the simulation approach in detail, Schützenmeister and Piepho first fitted the Cambridge data set with the above linear mixed model, then drew the QQ-plot of studentized conditional residuals and added 100(1-α)% simultaneous tolerance bands (STB) on the QQ-plot. Also, the studentized conditional residual vs. predicted values plot was drawn, and the 100(1-α)% simultaneous tolerance intervals (STI) were added on it. All the data points outside both the STB and STI were classified as outliers.

Specifically, the STB and STI were obtained following the procedure below:

1). Fit the data set with the linear mixed model

2). Use the fitted model to simulate 138 new response variables and refit the model with generated new responses

3). Repeat step 2) N times, which is known as the number of simulations

4). Save the N*138 studentized conditional residuals into a N-row, 138-column matrix, called "m".

Studentized conditional residuals were obtained by applying studentization on conditional residuals according to the equation below:

$$\hat{e}_k^* = \frac{\hat{e}_k}{\sqrt{\hat{p}_{kk}}}, \qquad (2)$$

where $\hat{p}_{kk}$ is an estimate of p_{kk}, the k-th diagonal element of matrix P.

$P = Var(\hat{e}) = RQR$, where $Q = V^{-1}(I - H)$, $V = Var(y) = ZGZ^T + R$, $H = XT$, and $T = (X^T V^{-1} X)^{-1} X^T V^{-1}$.

5). Sort all the elements in the 'm' matrix from minimum to maximum by rows and store them in an N*138 matrix, called 's'. Each row in the 's' matrix is an order statistic and we have $s_{j,1} \le s_{j,2} \le \ldots \le s_{j,138}$, for all j=1,2,…,N. After the s matrix is created, apply the following rank based deletion algorithm and delete a*N rows which contain the most extreme values (the minimum or the maximum studentized conditional residuals), and the remaining (1-α)*N rows of s matrix will form the STB.

- For each row in 's', locate the row(s) that contains the minimum studentized conditional residual(s) and delete that (these) row(s)
- Repeat the above step until α/2*N rows have been deleted, where a is the level of significance
- Delete the row(s) that contains the maximum studentized conditional residual(s)
- Repeat the above step until α/2*N rows have been deleted

6). After applying the deletion algorithm, follow the steps below to plot the STB on a figure with the remaining (1-α)*N rows in the s matrix

- Calculate the quantiles of each residual within each row of the s matrix. The plot below shows the details of the quantile calculation (SAS/STAT®(R) 9.2 User's Guide, Second Edition). F in the plot is the cumulative distribution function of a normal distribution.

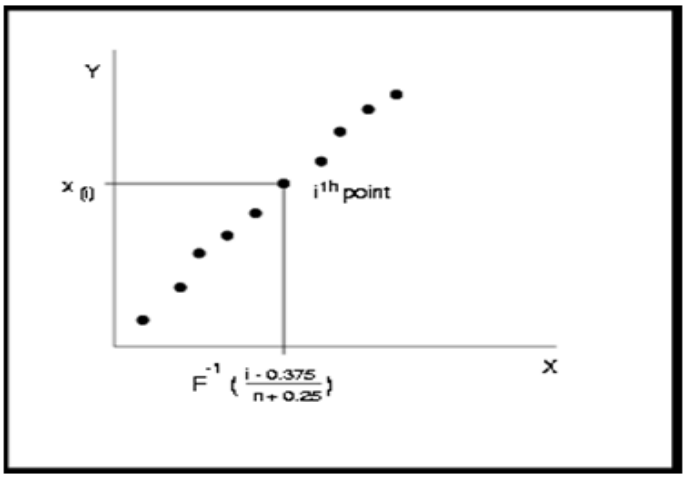

Figure 2.1 Construction of a Q-Q Plot

- Calculate the corresponding Z scores of each quantile and store in a vector, named “locationx”
- Plot each row of S against locationx

7) When the STB is plotted, add the ordered conditional studentized residuals on it, then the Q-Q plot is completed. All the conditional studentized residuals outside the STB are regarded as potential outliers.

8) While step 6 and 7 created a Q-Q plot with the STB and provided one of the two diagnosis criteria, this step is to create a residual plot with the STI. Plot the studentized conditional residuals vs. the predicted values and add the 95% STI horizontal lines, which are the minimum value of the 1st column in the s matrix and the maximum of the 138th column in the s matrix. Thus, the STI is created.

The above procedure is the detailed description on how the STB and STI plots were constructed and how to apply them. A series of comparisons and confirmation of this approach can be found in Appendix A.

The STI and STB would vary with different numbers of simulations, especially for STI, which only depends on two data points. It is therefore necessary to examine the number of simulations required for the STI and STB to reach convergence. Figure 2.2 describes the distribution of the upper and lower bound of the STI. The function of violin shape in the plot is the same as a histogram, and the wider part in a "violin" represents higher density. The distribution was obtained from 100 replications under each circumstance, from 10 simulations to 20000 simulations.

Figure 2.3 and 2.4 are distributions of upper and lower bound of the STB, respectively. Those distributions are also from 100 replications under each case, 10, 100, 1000, 10000 and 20000. Due to the limited space, not all the 138 data points' distributions of the STB are shown. Instead, only distributions of data points 1, 16, 31, 41, 61, 76, 91, 106, 121 and 138 are presented in figure 2.3 and 2.4. Combining distribution of the STI and STB plots, 10000 simulations are required for them to be stable.

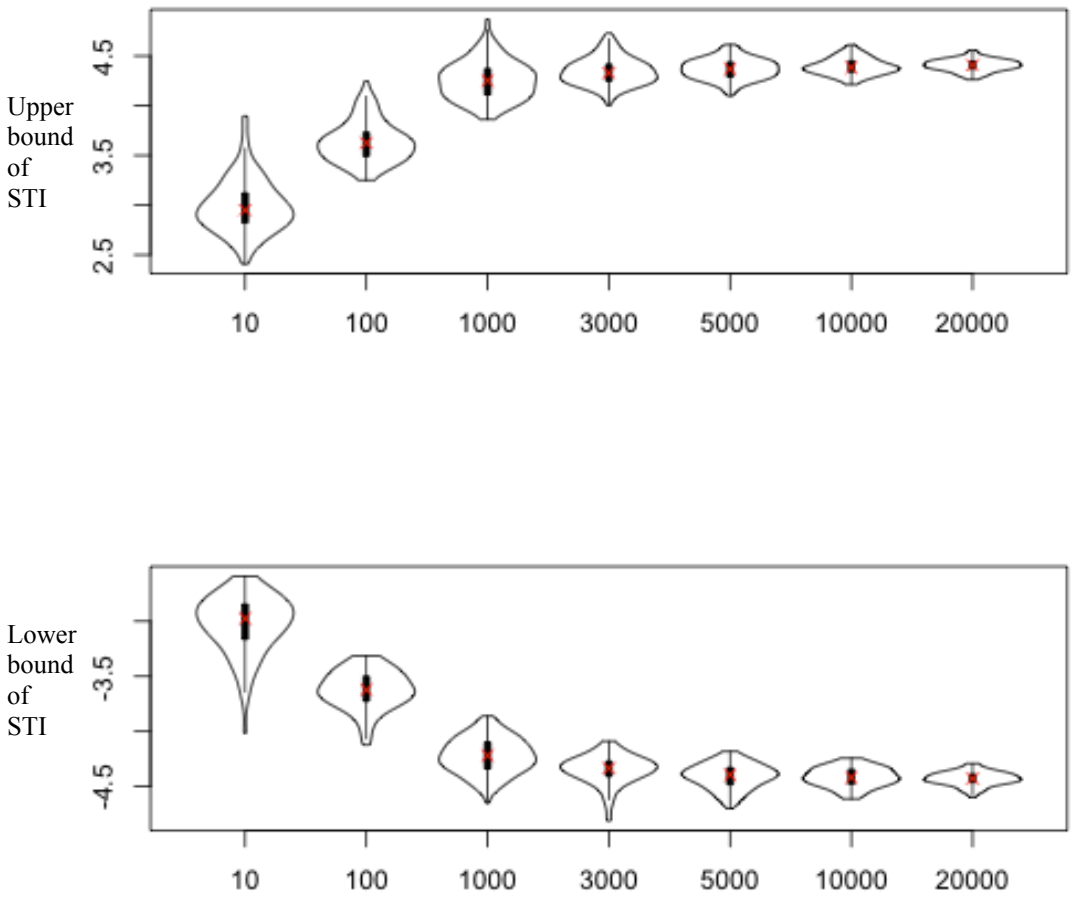

Figure 2.2 Distribution of upper and lower bound of the STI

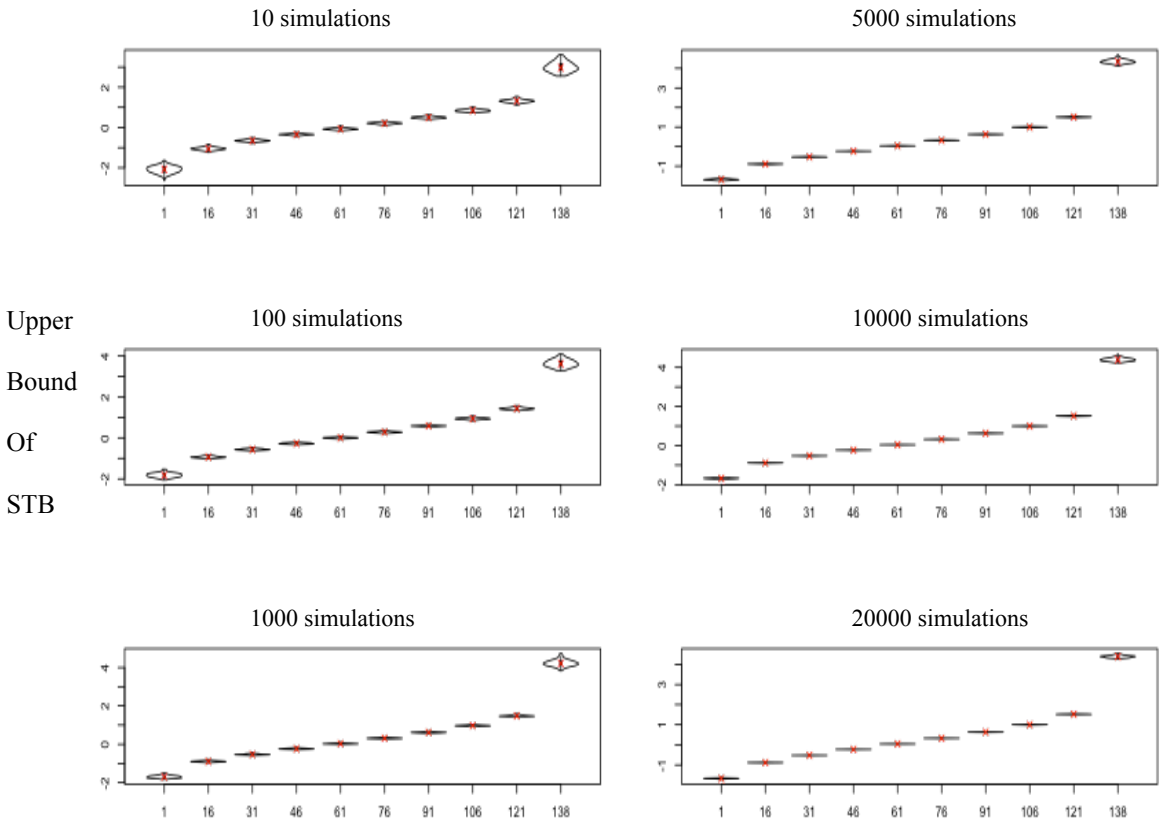

Figure 2.3 Distribution of upper bound of the STB

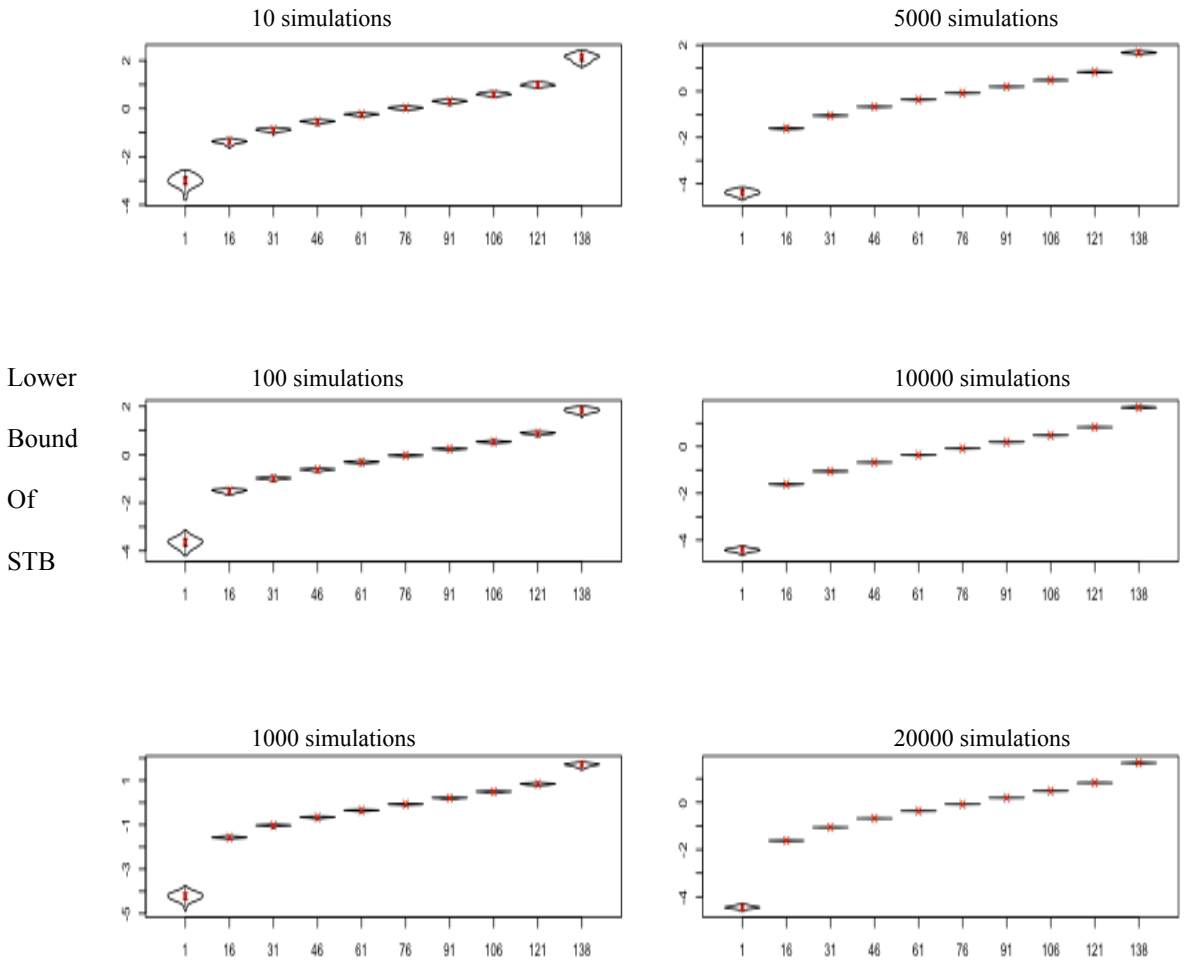

Figure 2.4 Distribution of lower bound of the STB

Chapter 3

EVALUATING THE PERFORMANCE OF THE SIMULATION APPROACH

To examine the limitation of this approach and to avoid the misuse of it, the performance of this simulation approach has been evaluated. This section includes two parts: (1) evaluating the capability of detecting "extreme outliers", and (2) evaluating the capability of detecting "true outliers".

"Extreme outliers" are data points that are obviously different from other points in the data set in terms of distribution. "True outliers" refer to the data points that are $3*\sqrt{\sigma_b{}^2+\sigma_e{}^2}$ away from the mean. They are not too obvious compared to the extreme outliers.

The evaluating process is actually to insert one or a few obvious outlier(s) in the normal data set and see how frequent this approach can detect it (them). The simulated Cambridge data set was used for this purpose. It is generated from simulation based on the original Cambridge model and the fitted parameters G and R.

3.1 Evaluating the Performance of Detecting Extreme Outliers

As shown in chapter 1, the model to fit Cambridge filter data is:

$$y_{ij} = \mu + a_i + b_j + e_{ij},$$

$$b_j \sim N(0,\sigma_b^2)$$

$$e_{ij} \sim N(0,\sigma_e^2)$$

or in the linear regression form as:

$$y = X\beta + Zb + e,$$

where X is the design matrix for the fixed effects and Z is the design matrix for the random effects. β is a vector of fixed effects and b is a vector of random effects. e is the random error term.

From both R and SAS analysis results, the fitted σ_b^2 is 0.001686* I_{14}, also known as G, and the fitted σ_e^2 is 0.000770* I_{138}, also known as R. The simulated y can be calculated via the following equation:

$$y_{sim} = X\beta + Zb_{sim} + e_{sim},$$

where b_{sim}~i.i.d. N (0,G)

e_{sim}~i.i.d. N (0,R)

From the simulation, we observe that the new response variables range from 0.0536 to 1.2971.

After the new data set simulated y is created, the model is refitted and new G and R parameters are generated, which are 0.00261492* I_{14} and 0.00079965* I_{138} respectively. The new G and new R are used in the following calculations, which is quite similar to the procedures described in Chapter 1.

(1) Calculate the P matrix, which equals to $Var(\hat{e}) = RQR$,

where $Q = V^{-1}(I - H)$, $V = Var(y) = ZGZ^T + R$, $H = XT$, and

$$T = (X^T V^{-1} X)^{-1} X^T V^{-1}$$

(2) Calculate the predicted y, according to:

$$y_{predicted} = X\hat{\beta} + Z\hat{b},$$

where $\hat{\beta} = Ty_{sim}$ and $\hat{b} = GZ^T V^{-1}(y_{sim} - X\hat{\beta})$

(3) Calculate the conditional residuals by subtracting $y_{predicted}$ from y_{sim}.

(4) Studentize the conditional residuals according to:

$\hat{e}^*_k = \frac{\hat{e}_k}{\sqrt{\hat{p}_{kk}}}$, where $\hat{p}_{kk}$ is the k-th diagonal element of matrix P.

(5) Create S matrix and conduct deletion algorithm.

(6) Plot the 95% STI and STB

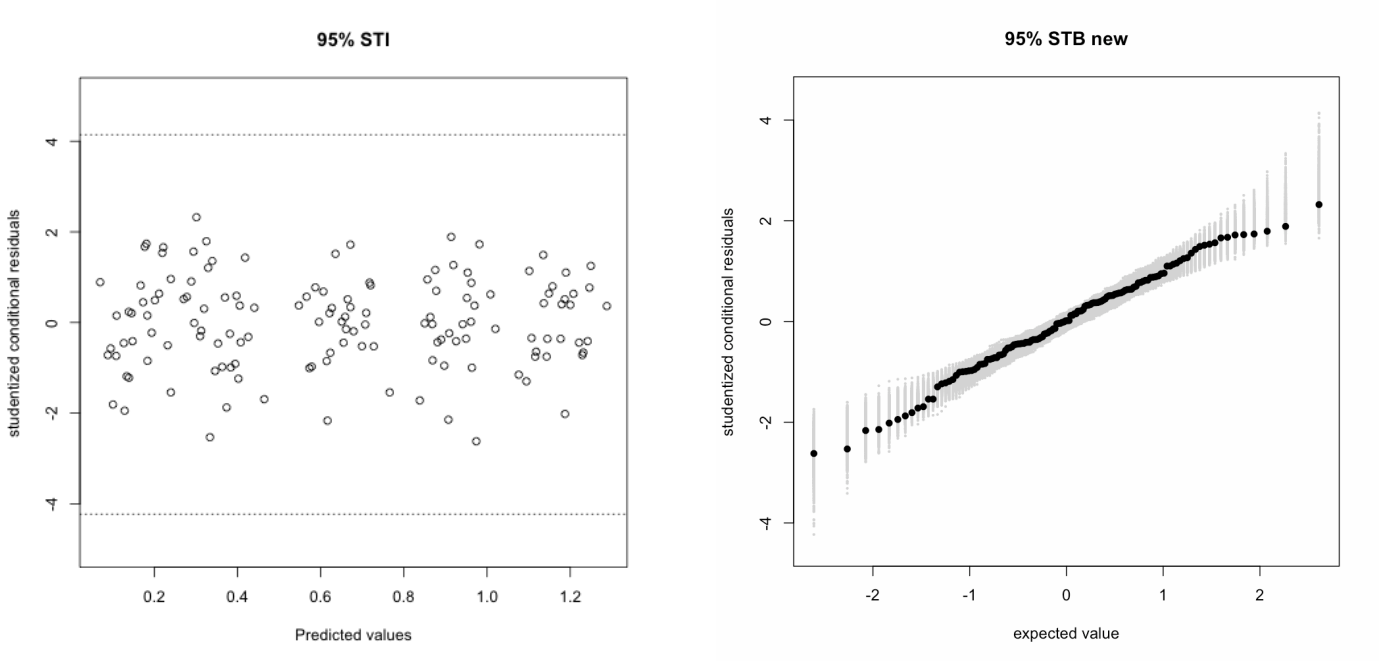

Figure 3.1 The 95% STI (left) and STB (right) for the simulated Cambridge data

The above plots show that all the simulated response variables, which are i.i.d. normally distributed, are located within the two bands. When one point in simulated y is changed to an obvious outlier, the STB plot changes dramatically.

One example is when the 50^{th} simulated y changes from 1.1356 to 3, the corresponding STB and STI look as below:

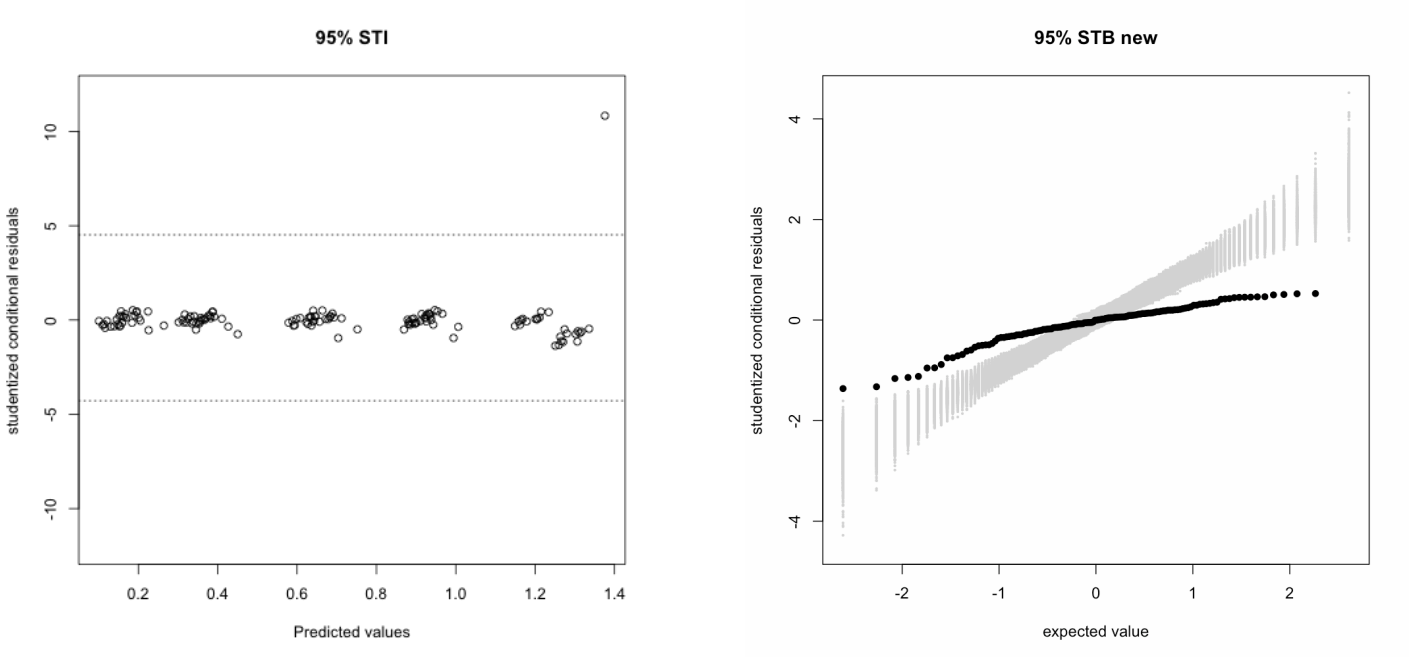

Figure 3.2 The STI (left) and STB (right) when the 50th simulated y changed to 3

As is shown in the above two plots, the STI indeed detected the 50^{th} simulated y as an outlier. However, there are too many points outside the STB, since the shape of the distribution of the data points in the STB plot changes greatly.

The reason for this to happen is that even one obvious outlier point can dramatically change the fitted parameters, resulting in a significant change in the model. Thus, all the residuals are changed, and the locations of these points in the STB plot are all shifted. Since only one outlier will result in the significant change of the shape of the normal QQ plot, we cannot only rely on the STB plot to make judgment. Instead, the result of the STI should be combined with the STB result to tell whether a data point is an outlier or not. Only when a point is outside both the STB and STI, it can be regarded as an outlier. If it is merely outside the STB, the evidence is not sufficient to classify it as an outlier.

Refer to the rule of identifying the residuals, "the points outside of both the STB and STI are regarded as residuals", this method detects the obvious outlier successfully at this time.

However, it does not happen all the time when there is more than one outlier. For example, when the 50^{th} simulated y changes to 0.02 and 70^{th} simulated y changes from 0.2279 to 0.02, it only detects the former point as shown in figure 3.3.

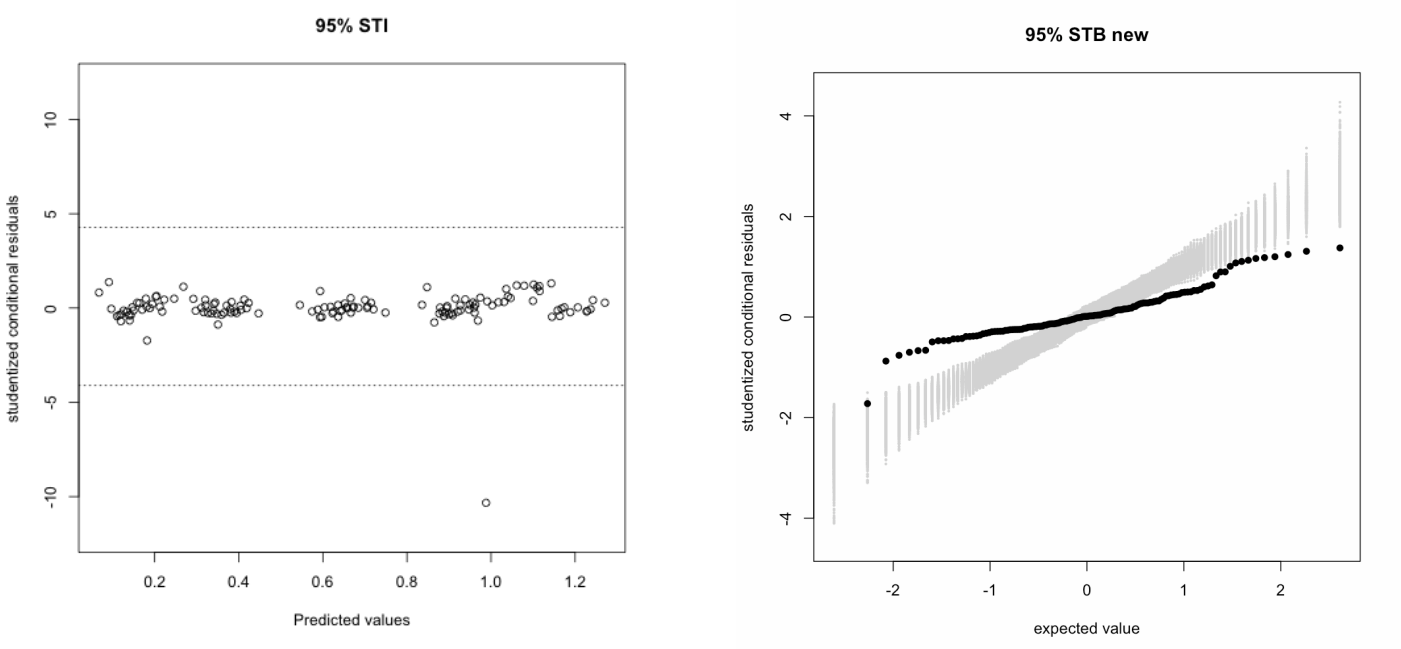

Figure 3.3 The STI (left) and STB (right) when the 50th simulated y changed to 0.02 and 70th simulated y changed to 0.02

Different combinations of changed points have been used to test the performance of this approach. A summary table is provided to summarize the testing results.

Table 3.1 Summary table for simulated Cambridge data testing result

Data range (0.0536-1.2971)	Change in y	The STI detected or not	New G	New R
G: 0.00261492 R: 0.00079965	Y50->5	Yes	2.7485e-12* I_{14}	0.11012* I_{138}
	Y50->3	Yes	0.0024183* I_{14}	0.0255921* I_{138}
	Y50->2	Yes	0.0025236* I_{14}	0.0060591* I_{138}
	Y50->1.5	Yes	0.0025762* I_{14}	0.0017015* I_{138}
	Y50->2 Y70->1.5	Yes	0.0028887* I_{14}	0.0183005* I_{138}
	Y50->3 Y70->1.5	Yes	0.0026145* I_{14}	0.0379986* I_{138}
	Y50->0.02 Y70->0.03	Only Y50 was detected	0.0026274* I_{14}	0.0102201* I_{138}
	Y50->0.02 Y130->0.02	Only Y50 was detected	0.002622* I_{14}	0.010248* I_{138}
	Y70->0.02 Y130->0.02	Yes	0.0023627* I_{14}	0.0013798* I_{138}

It can be seen in the above table that when there is only one outlier, G is decreasing while R is increasing when a certain point changes to a more extreme outlier. For example, when the 50th simulated y changes to 2, G becomes 0.0025* I_{14}

and R becomes 0.006* I_{138}. When the same data points changes to 3, G decreases a little bit and becomes 0.0024* I_{14} and R increases to 0.026* I_{138}. When it changes to 5, G drops greatly to 2.75e-12* I_{14} and R increased to 0.11* I_{138}. Another thing worth mentioning is that when there is only one extreme value, the successful detection rate of this method is very high.

When there are two outliers, similar pattern of the change in G and R can be found. For example, when simulated y70 is changes to 1.5, G is smaller and R is bigger when simulated y50 is 3 compared to simulated y50 equals to 2, which is a less obvious outlier. However, whether or not this method can detect both of the outliers is not easy to determine. It depends on how extreme the new value is and how different the new value is from the original value. For example, when y50 changes to 2 and y70 changes to 1.5, this method detects both outliers successfully. It is also the case when y50 changes to 3 and y70 changes to 1.5. However, this method fails when y50 changes to 0.02 and y70 changes to 0.03 or y130 changes to 0.02. In both cases, it only detects y50. This is because y50 changes from 1.136 to 0.02, which is a significant change. While for y70 or y130, they change from 0.228 or from 0.200 to 0.02, which is less significant. When both y70 and y130 change to 0.02, this method works again. It is because they change from similar values to the same value, which is a more balanced variation. In a word, whether or not this method works depends on a lot of factors such as the number of extreme outliers and difference between the new outliers and the original values.

All the above testing results are based on replacing data points with obvious outliers. Actually, those extreme outliers can be separated apart from the remaining

data set easily by visual inspection. And to detect those obvious outliers, we do not even need to depend on a certain approach or algorithm.

Considering the significant impact of extreme values on the performance of this approach, we suggest the obvious outliers are eliminated before analyzing the real data set with this approach.

3.2 Evaluating the Performance of Detecting True Outliers

Begin with the same simulated Cambridge data set obtained in Section 3.1, one, two or three true outliers were inserted into the data set. The true outliers are generated from the steps below:

1. One, two or three data point(s) are randomly selected from the simulated Cambridge data set.
2. Add $6*\sqrt{\sigma_b{}^2+\sigma_e{}^2}$ to this (these) selected value(s) and update the response variable data set, where $\sigma_b{}^2 = 0.00261492$ and $\sigma_e{}^2 = 0.00079965$.They are obtained from the model fitting result with the simulated Cambridge data set.
3. Apply simulation approach and record how many times those outliers are detected.

The reason for adding $6*\sqrt{\sigma_b{}^2+\sigma_e{}^2}$ s that since true outliers are referred the values that are $3*\sqrt{\sigma_b{}^2+\sigma_e{}^2}$ away from the mean, and there are some data points in the simulated Cambridge data set close to $-3*\sqrt{\sigma_b{}^2+\sigma_e{}^2}$ due to the nature of simulation, adding $6*\sqrt{\sigma_b{}^2+\sigma_e{}^2}$ o any randomly selected numbers will guarantee that the new values would be at least $3*\sqrt{\sigma_b{}^2+\sigma_e{}^2}$ away from the mean. From the preliminary testing (results not shown), when adding the number $s*\sqrt{\sigma_b{}^2+\sigma_e{}^2}$, where s ranges from 3 to 6, to the randomly selected data points,

the successful detection rate increases with increased s and achieves 100% when s equals to 6.

Below is a summary table for the successful detection rate when s is 6.

Table 3.2 Successful detection rates out of 100 trials

	One outlier	Two outliers	Three outliers
Successful detection rate	100%	100%	94%

The above table shows that when there are one or two "true outliers", this simulation approach can detect the outliers successfully. When there are three outliers, however, the detection fails 6 times out of 100. Each time, two outliers of those 6 unsuccessful tests are identified, with only one outlier undetected. The possible reason for one outliers out of three being unidentified could be that this data point was -3 standard deviations from the mean while the other two points were 3 standard deviations from the mean. After adding 6 standard deviations to the three numbers, one point becomes 3 standard deviations from the mean and the other two points are 9 standard deviations from the mean. Therefore, the importance of the first outlier is "masked" by the other two outliers and thus unidentified.

This table discloses another limitation of this approach - i.e., when the distances of the few outliers in the data set from the mean are different to a certain extend (e.g. 6 standard deviations), this approach sometimes only identifies the more extreme outliers. On solution could be that we apply this approach more than once. First, apply this approach to delete the identified outliers. Then, apply it again to check if there are more outliers.

Chapter 4

ANALYSIS OF REAL DATA SET

In the previous chapters, this method was used to analyze relatively simple data sets with simple mixed models. In this chapter, a more complex data set will be analyzed to evaluate the capability of this approach in dealing with a more complicated model.

4.1 Data Description

The data set used here is the yield data from a 2011 state trial. The explanatory variables in this data set include: environment factor, block, entry, maturity group and moisture level in grain. The response variable is the yield. The histogram of this data set is shown as below:

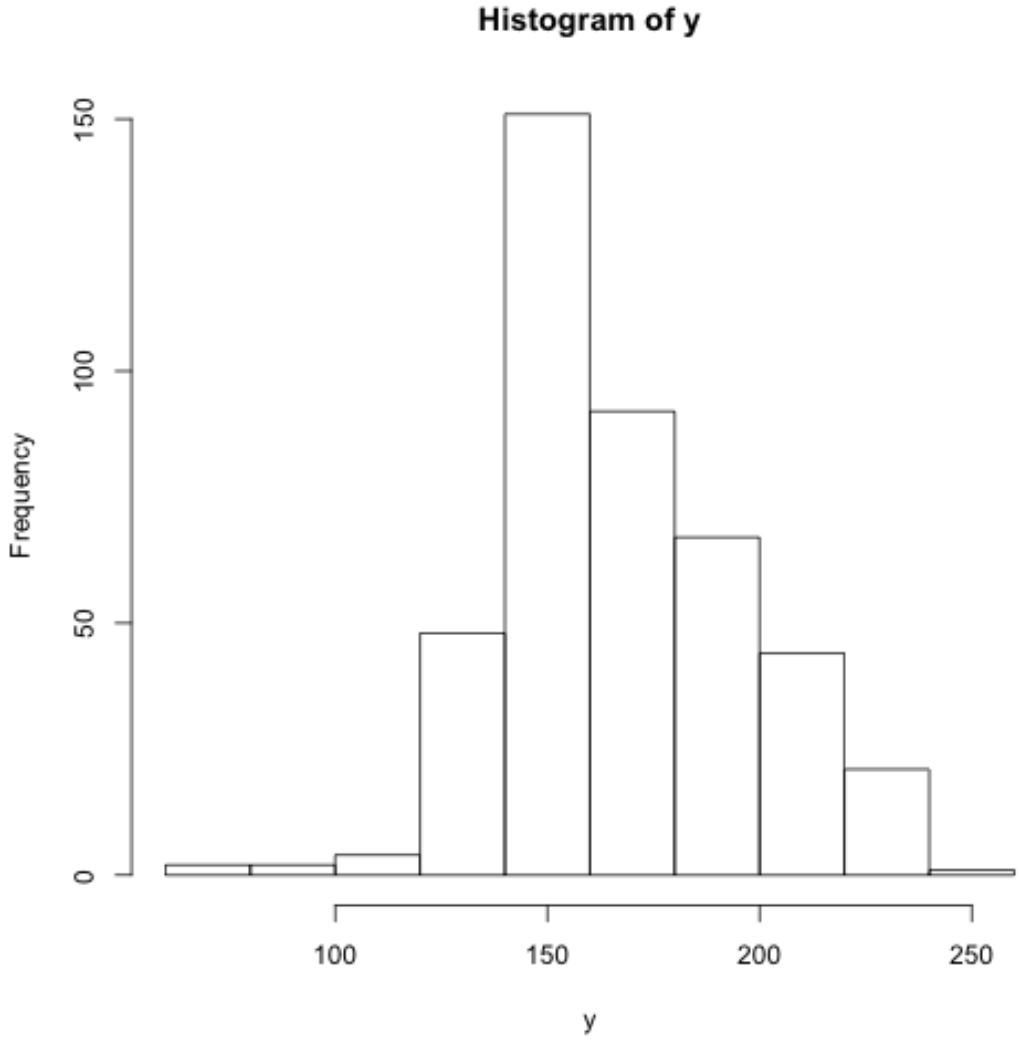

Figure 4.1 Histogram of the yield data

4.2 Model Fitting

The model to fit the data is:

$$yield_{ijkg} = u + maturity_i + entry_j + environment_k + environment_k * maturity_i + block_g(environment_k * maturity_i),$$

where i=1,2,3; j=1,2,3,…,23; k=1,2; g=1,2,3,4.

Among those variables, environment and block are random variables. There are 432 observations in total.

Since this is a more complex model, visualizing the design matrix is more difficult. A R package "RLRsim" is used to extract the design matrices for this model. The procedure to generate the STB and STI is the same except for the design matrices extraction step. The procedure includes:

(1) Fit the model with R

(2) Extract the design matrices with RLRsim package

(3) Calculate conditional procedure and go through studentization

(4) Apply the deletion algorithm and create the STI and STB

4.3 Analysis Results

The following table is a summary of fitting results of the model. The variances for block, interaction between environment and maturity and environment factors are 45.89, 544.70 and 156.53 respectively. Residual variance is 365.43. Among the fixed effects, maturity group 2 has the highest estimated effect on the yield, which is 168.50, whereas maturity group 3 is with the lowest estimated effect on the yield. The three entries with highest estimated effects are entry 6, entry 5 and entry 14. The estimated

effect of entry 15 is only 0.629 smaller than that of 14, as a result, entry 15 is ranked the 4th highest. The three entries have the lowest estimated effects are: entry 20, entry 2 and entry 21.

Table 4.1 Summary table for the yield data analysis result

Random effects estimates	Variance	Standard Deviation
Block(Maturity*Environment)	45.893	6.7745
Maturity*Environment	544.696	23.3387
Environment	156.531	12.5112
Residual	365.427	19.1161

Fixed effects estimates	Estimate	Std. Error	t value
Mat1	160.071	19.332	8.280
Mat2	168.500	19.319	8.722
Mat3	148.167	19.319	7.669
Entry2	-0.875	5.518	-0.159
Entry3	3.237	5.518	0.587
Entry4	9.854	5.518	1.786
Entry5	15.871	5.518	2.876
Entry6	17.950	5.518	3.253
Entry7	6.871	5.518	1.245
Entry8	11.329	5.518	2.053
Entry9	10.717	5.518	1.942
Entry10	6.429	5.518	1.165
Entry11	14.067	5.518	2.549
Entry12	5.817	5.518	1.054
Entry13	5.804	5.518	1.052
Entry14	14.410	6.217	2.318
Entry15	10.923	6.217	1.757
Entry16	7.423	6.217	1.194
Entry17	2.785	6.217	0.448
Entry18	4.504	6.217	0.724
Entry19	6.712	7.922	0.847
Entry20	-2.776	7.922	-0.350
Entry21	-0.588	7.922	-0.074
Entry22	5.300	7.922	0.669
Entry23	10.225	7.922	1.291

The STI and STB plots based on 1000 simulations are shown in Figure 4.2 and Figure 4.2. And the detected outliers by the STI and STB are points 392, 393, 409 and 411. All of them come from Environment 2, Maturity group 2 and block 4. They are entry 1, 2, 18 and 20. The range of this data is from 66.9 to 246.5. Point 393 happened to be the smallest number: 66.5 and values of the rest of these detected outliers are: 72.5, 80.7 and 93.6. These detected outliers are the smallest 4 points in the data set. Therefore, it is valid to consider them as outliers.

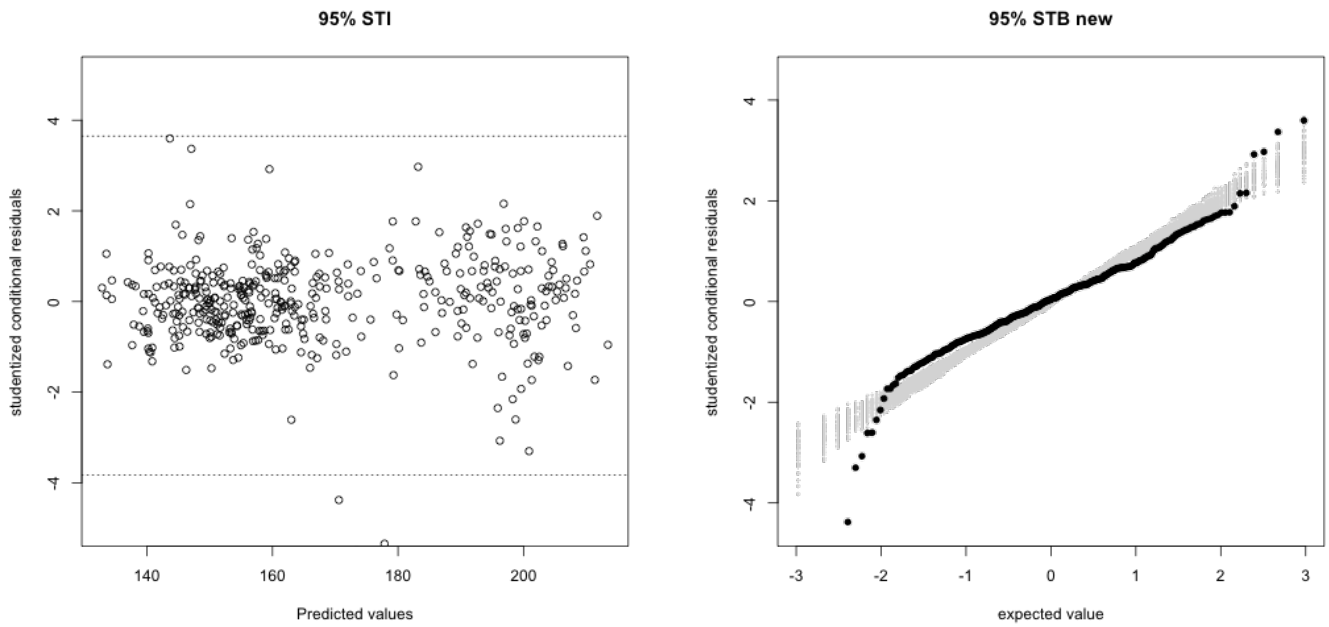

Figure 4.2 The STI (left) and STB (right) of the yield data set

The STI and STB plots based on 10000 simulations are also created. However, because of their large sizes, they are not shown here. When the simulation number increased from 1000 to 10000, both the STI plot and STB plot become wider. As a result, in the first trial of this approach on the data set, only three data points are identified as outliers, which are the point 392, 393 and 409. Then the detected three outliers in the first step are deleted and the second trial has applied on the remaining data set. In the second trial the data point outstanding both the STI and STB is the point 411. Then this point is deleted and the third trial has applied, the result shows no outliers this time.

In a sum, when the number of simulations is 1000 or 10000 the same four points are found as outliers. As the number of simulations increases, this approach tends to be more conservative in the first trial. However, after more than one trial have been applied on the data set, the same result is likely to achieve.

DISCUSSION

A linear mixed model is a natural extension of the linear model. It has many advantages over the linear model and is widely used in agriculture science such as genomic selection in animal and plant breeding. Outlier detection is an important step in data analysis of linear mixed models. Despite the existence of formal outlier tests in linear models, there is no easy approach for outlier identification in linear mixed models. Analyzing outliers in linear mixed models is usually with the aid of the normal QQ plot and the residual plot. However, it inevitably involves a certain level of subjectivity. Schutzenmeister and Piepho proposed a simulation approach, that is, to add the 100(1-α)% STB and STI to normal QQ plot and residual plot. All the data points outside both the STB and STI are regarded as potential outliers. This method objectifies the interpretation of the analysis results.

Based on Schutzenmeister and Piepho's approach, an open-sourced statistical tool is developed to automatically identify potential outliers. This tool is convenient to use and can be made accessible to smallholder farmers in developing countries. It may also help inexperienced analysts obtain useful information with their limited time and budgets.

The steps of the procedure of this approach includes: fit the data set to analyze with an appropriate model, simulate new response values and refit the model, obtain studentized conditional residuals, apply deletion algorithm, create the STI and STB and draw conclusions at the end. The correctness of the R code to realize each step is confirmed.

The distribution of the STI and STB was obtained by repeating the simulation steps 100 times. The results indicate that 10000 times of repetition is sufficient for simulation.

Also, the performance of this approach has been evaluated, which discloses two disadvantages of the method. Firstly, it sometimes fails to detect extreme outliers, since the G and R parameters change dramatically with only one obvious outlier in data set. Due to the unstable performance of this approach in detecting obvious outliers, we suggest eliminate any extreme data from the data set before proceeding to use this method. Another disadvantage is that when there are multiple true outliers in the data set, the approach sometimes only detects the more extreme outliers and leaves the less obvious outliers unidentified. The solution could be to apply the method once, and identify a few outliers first, then discard those points and repeat this approach again and check if more outliers can be identified.

To further apply this approach, one complex model with multiple random variables is used. The method successfully detects 4 potential outliers from the real data set and the result is satisfactory.

REFERENCES

Gbur, E. E., 2012. Analysis of Generalized Linear Mixed Models in the Agricultural and Natural Resources Sciences. American Society of Agronomy, Soil Science Society of America, Crop Science Society of America, Madison, WI.

Andrade-Bejarano, M., Longford, N.T., 2010. Outliers in mixed models for monthly average temperatures. Austrian Journal of Statistics 39, 203-221.

Gumedze, F. N., Welham, S.J., Gogel, B.J., Thompson, R., 2010. A variance shift model for detection of outliers in the linear mixed model. Computational Statistics and Data Analysis 54, 2128-2144.

Kutner, M. A., Nachtsheim, C., Neter J., 2004. Applied linear Regression Models. 4th ed. McGraw-Hill Publishing Company, Colunbus, OH.

Littell, R. C., 2006. SAS for Mixed Models. 2nd ed. SAS Institute, Inc., Car, NC.

Longford, N. T., 2001. Simulation-based diagnostics in random-coefficient models. J. R. Statist. Soc., 164, 259-273.

Nobre, J.S., Singer, J.M., 2007. Residual analysis for linear mixed models. Biometric Journal 49, 863-875.

Pollock, L., Wisser, R. J., Meyers, B., Lee, J. S., Smith, M., 2011. BREAD: Automating and Integrating Processes to Improve Plant Breeding for Smallholder Farmers.

Schützenmeister, A., Piepho, H., 2012. Residual analysis of linear mixed models using a simulation approach. Computational Statistics and Data Analysis 56, 1405-1416.

Appendix

COMFIRMATION OF THE CORRECTNESS OF THE SIMULATION APPROACH

In the thesis, a detailed description on how the STB and STI plots are constructed and how to apply them are discussed. A series of comparison and confirmation of this approach are presented in this appendix. The data set used for the purpose of comparison is the Cambridge filter data. Confirmation of model fitting, simulation and calculation with R code are achieved by comparing R results and SAS results. Confirmation of rank based deletion algorithm is through plotting the STBs on the same plot with R code and Schützenmeister and Piepho's R code. Comparison of analysis results is accomplished by inspecting the final STI and STB plots obtained by R code and the figures available in Schützenmeister and Piepho's paper.

A.1 Confirmation of the R Code with SAS Result

Before proceeding to rank-based deletion algorithm, there are some data processing steps: the model has to be fit, the 'm' matrix has to be created through simulations and studentization has to be applied on the conditional residuals. This section is to confirm whether those three steps in R code are correct or not.

A.1.1 Model Fitting-Fixed and Random Effects by R & SAS

Model fitting generates fixed-effects parameters and random-effects parameters. The table below shows model fitting results from R and SAS. Lme4 package in R is used in the model fitting process and Mixed Procedure is used in SAS.

As shown in Table 1, all the estimated 10 levels of fixed effects by R are the same as that by SAS. Also, all of them have the same standard errors and t values. In addition, Table 1 shows the consistence between the random effects estimates by R and SAS. Parameter G from both methods is 0.001686 and parameter R from both methods is 0.000770.

Table A.1 Model fitting comparison

		Results from R			Results from SAS		
Random effects estimates	Lab	0.00168611			0.001686		
	Residual	0.00077025			0.000770		
Fixed effects estimates		Estimate	Std. Error	t value	Estimate	Std. Error	t value
	Sample1	0.16250	0.01325	12.27	0.1625	0.01325	12.27
	Sample2	0.18429	0.01325	13.91	0.1843	0.01325	13.91
	Sample3	0.35886	0.01325	27.09	0.3589	0.01325	27.09
	Sample4	0.39907	0.01325	30.13	0.3991	0.01325	30.13
	Sample5	0.64464	0.01325	48.67	0.6446	0.01325	48.67
	Sample6	0.67036	0.01325	50.61	0.6704	0.01325	50.61
	Sample7	0.95150	0.01325	71.83	0.9515	0.01325	71.83
	Sample8	0.91336	0.01325	68.95	0.9134	0.01325	68.95
	Sample9	1.21805	0.01342	90.74	1.2181	0.01342	90.74
	Sample10	1.16244	0.01342	86.60	1.1624	0.01342	86.60

A.1.2 Comparison of Simulation Results

The ‘m’ matrix is obtained by simulating new responses according to the fitted model, followed by refitting the model with the simulated new response. The effectiveness of this step in R code can be confirmed by comparing the refitted model parameters from R and SAS. Firstly, new responses are simulated with R. Then the same set of simulated data is used for refitting the model by R and SAS respectively.

According to Table 2, the refitted models by R and SAS have the same estimated fixed effects and random effects. Parameter G from both methods is 0.001263 and parameter R from both methods is 0.000797. Ten levels of fixed effects also have the same estimates, standard errors and t values.

Table A.2 Comparison of simulation results

		Results from R			Results from SAS		
Random effects estimates	Lab	0.00126343			0.001263		
	Residual	0.00079682			0.000797		
Fixed effects estimates		Estimate	Std. Error	t value	Estimate	Std. Error	t value
	Sample1	0.15771	0.01213	13.00	0.1577	0.01213	13.00
	Sample2	0.17665	0.01213	14.56	0.1767	0.01213	14.56
	Sample3	0.34862	0.01213	28.74	0.3486	0.01213	28.74
	Sample4	0.39270	0.01213	32.37	0.3927	0.01213	32.37
	Sample5	0.63849	0.01213	52.64	0.6385	0.01213	52.64
	Sample6	0.65707	0.01213	54.17	0.6571	0.01213	54.17
	Sample7	0.93710	0.01213	77.26	0.9371	0.01213	77.26
	Sample8	0.90736	0.01213	74.81	0.9074	0.01213	74.81
	Sample9	1.21045	0.01233	98.17	1.2104	0.01233	98.17
	Sample10	1.16923	0.01233	94.83	1.1692	0.01233	94.83

A.1.3 Comparison of Calculations

Another key step in R code is to calculate predicted response value, conditional residuals and studentized conditional residuals, for construction of the STB and STI and residual diagnostics are all based on conditional studentized residuals. Studentization process is a relatively complex step, for it involves calculating matrix "P", which equals to "RQR", Hat matrix, Design matrixes and variance of y. Table 3 indicates there are slight differences between the results from R and SAS. For

predicted y values, R and SAS results are the same until their millionths decimal places. For conditional residuals, they also begin to differ after their millionths decimal places. However, after studentization, they began to differ from ten-thousandths decimal places. These differences are within an acceptable range, and could be resulted from aggregations in error.

Table A.3 Comparisons of predicted response value, conditional residuals and studentized conditional residuals

Predicted Y		Conditional Residuals		Studentized Residuals	
From SAS	From R	From SAS	From R	From SAS	From R
0.1794206099	0.17942082	-0.01842061	-0.01842082	-0.72493511	-0.724462149
0.2012063242	0.20120654	-0.009206324	-0.009206539	-0.362310893	-0.362078743
0.3757777527	0.37577797	-0.002777753	-0.002777968	-0.109317254	-0.109253111
0.4159920385	0.41599225	0.0010079615	0.001007747	0.0396678892	0.039633095
0.661563467	0.66156368	0.001436533	0.001436318	0.0565341318	0.056488137
0.6872777527	0.68727797	0.0047222473	0.004722032	0.1858419919	0.185710127
0.9684206099	0.96842082	-0.01642061	-0.01642082	-0.646225978	-0.645805286
0.9302777527	0.93027797	-0.012277753	-0.01227797	-0.483185632	-0.482873212
1.2349746009	1.23497479	0.0530253991	0.05302521	2.0928326718	2.091440941

Through the above 3 steps of confirmation, we are confident that the 'm'matrix has been correctly built

A.2 Confirmation of the Correctness of Rank-based Deletion Algorithm

After 'm' matrix is created, all the values in each row of m will be ordered from smallest to largest and each ordered statistics would become a row in s matrix. A rank based deletion algorithm is then to apply on 's'. It is also an essential step in that the remaining rows in the 's' matrix after deletion would form the STB and STI. The

R code segment of this deletion algorithm is obtained from Schützenmeister and Piepho. Beginning with the same s matrix, deletion algorithms of the R code in this work and Schützenmeister and Piepho's R code are applied respectively. The STB is obtained by the R code in this work and Schützenmeister and Piepho's R code are plotted in the figure (Figure 1) below. The former STB is in light grey color while the latter is in red. This figure shows the red points are in a good match with the boundaries of the light-grey bands. It verifies the deletion approach in the R code is consistent with Schützenmeister and Piepho's algorithm.

Figure A.1 Confirmation of deletion algorithm: red points are obtained by Schützenmeister and Piepho's R code, light grey bands are obtained by R code, black squares are conditional studentized residuals.

A.3 Comparison of Analysis Results

Following deletion algorithm is the analysis result. The outlier diagnosis is composed of two parts: locate all the potential outliers by the STB and find out all the potential outliers by the STI. If the data points are outside both the STB and STI, they will then be regarded as outliers. Since the same data set is used, and the confirmed R code is adopted, it's reasonable to assume that the diagnosis results should match that available in Schützenmeister and Piepho's paper. The following sections compare the STB results and the STI results respectively.

A.3.1 Comparison of Analysis Results from the STB

As shown in Figure 2, these two QQ-plots of studentized conditional residuals of Cambridge data sets look quite similar, with the same shape and trend. Table 4 is a summary of all the outlying points in figure 2 (a). It shows that points 129, 130, 118 and 31 are outside the STB in (a), which is in consistence with the result in Figure 2 (b). Also, as is shown in Figure 2 (b), points 137, 138 and 117 are within the bands, and these points are not outlying points in figure 2 (a). Both Figure 2 and Table 4 provide evidence that STB analysis is consistent with that of Schützenmeister and Piepho's result.

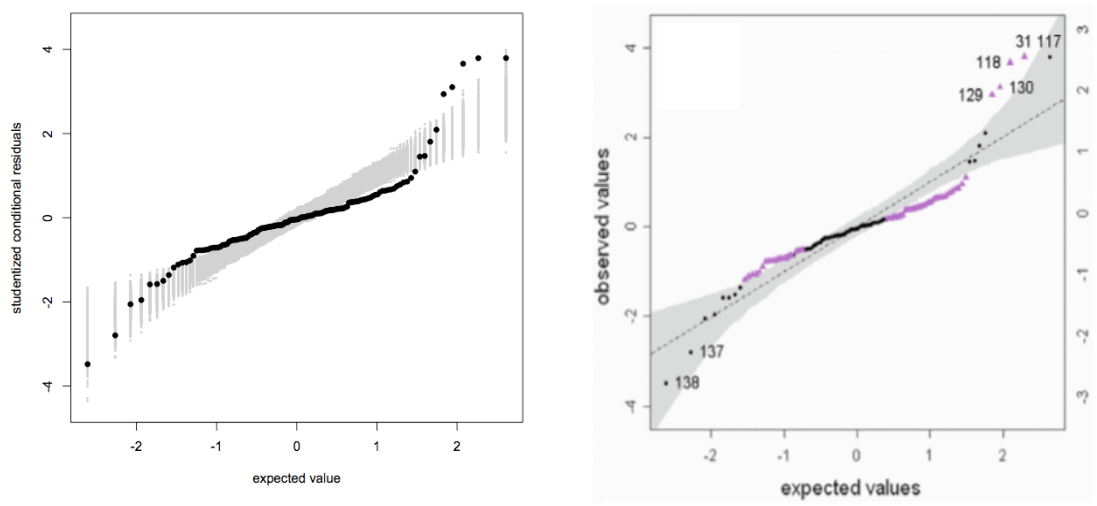

Figure A.2 QQ-plot of studentized conditional residuals of Cambridge filter data: (a) Left is generated by the R code and (b) the right is from Schützenmeister and Piepho's paper

Table A.4 Detected outlying points by the R code

```
> outliers
  [1]  29 132 127  11 113  97   1  84 128  77  51   7  30
 [14] 112  78  28  26 120  16 124  82  89  27  96   6  65
 [27] 131  49  18 121  74 122  10  24  81  67  75  92  80
 [40]  40  22  33 104  88  87  91  70  56  38  23  93  45
 [53]  48  83  72  57 108  63  46 129 130 118  31   0   0
 [66]   0   0   0   0   0   0   0   0   0   0   0   0   0
 [79]   0   0   0   0   0   0   0   0   0   0   0   0   0
 [92]   0   0   0   0   0   0   0   0   0   0   0   0   0
[105]   0   0   0   0   0   0   0   0   0   0   0   0   0
[118]   0   0   0   0   0   0   0   0   0   0   0   0   0
[131]   0   0   0   0   0   0   0   0
```

A.3.2 Comparison of Analysis Results from the STI

In Figure 3 (a), the upper bound of the STI is 3.989 and the lower bound is -4.350. There is no data point outside this interval. Since only if a point is outside both the STB and STI, it is regarded as an outlier, the R code fails to detect any outliers. However, Schützenmeister and Piepho's interval is narrower (Figure 3 (a)), with a value between 3 and 4 as the upper bound and a value between -3 and -4 as the lower bound. Points 31, 117, 118 and 138 are outside this interval. Referring to Figure 2 (b), the only data points outside both the STB and STI is point 118 and it is classified as outliers by Schützenmeister and Piepho.

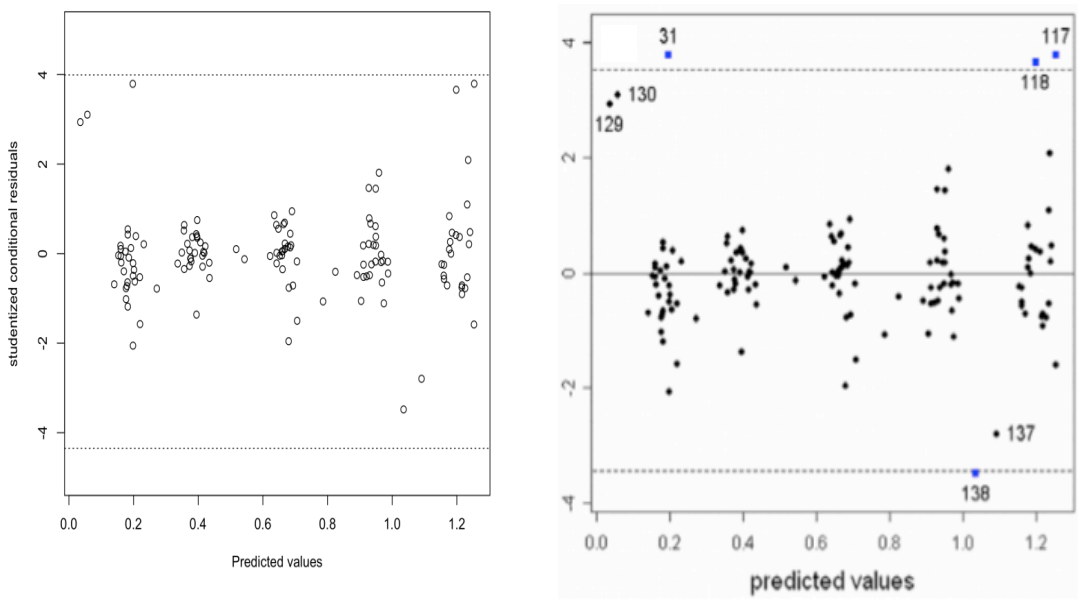

Figure A.3 Residual-plot of studentized conditional residuals of Cambridge filter data: (a) Left is generated by the R code and (b) the right is from Schützenmeister and Piepho's paper

Actually, there are some details not valid in Schützenmeister and Piepho' figure, unless they make some adjustment of their STI, which is not mentioned in their paper. Recalling that the STB is formed by all the ordered statistics in S matrix after the deletion algorithm, if we draw all the number as light-grey points in a QQ plot by rows in 's', all the points that have the same x-axis position are from the same column of S. Therefore, it is obvious to identify from the STB plot which is the minimum number in the first column of S-the lowest light grey points at the very left of the STB. Also, the maximum number in the last column of S should be the highest light grey points at the right margin of the STB. Those two points happen to be the lower and upper bound of the STI, respectively. Therefore, there should be a match in the STB and STI, as shown in Figure 4. The STB plot indicates the minimum value of the first column of s matrix should be a value smaller than -4, however, the lower bound of the STI is between -4 and -3. In addition, the maximum value of the last column of s matrix should be a value greater than 4 according to the STB plot. But the STI's upper bound is obviously smaller than 4. That is the reason why it is so narrow and can successfully detect some outliers. The contradiction in these two plots reveals possible mistakes in Schützenmeister and Piepho's paper. An approach to fix them should exist. Otherwise, no outliers will be detected according to their method.

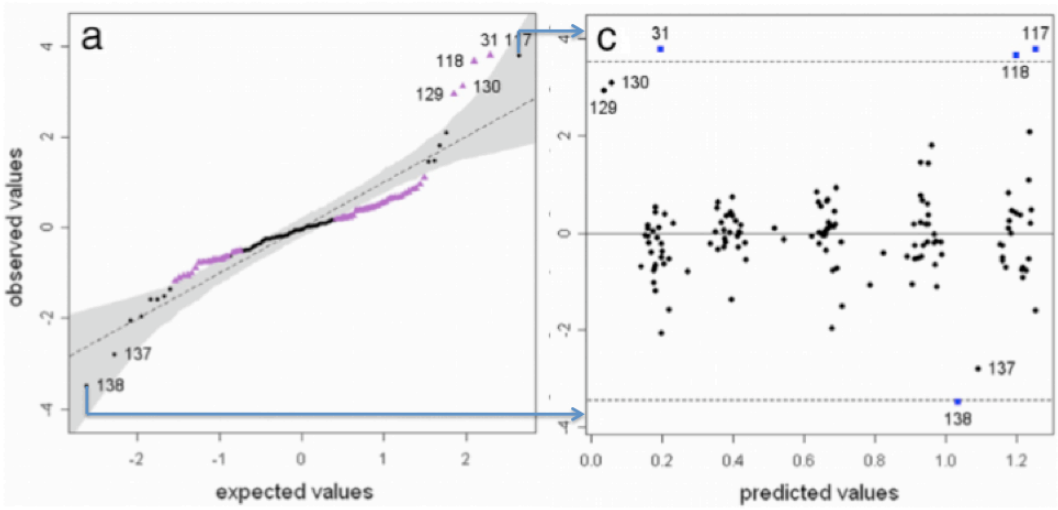

Figure A.4 Contradiction in the STB and STI in Schützenmeister and Piepho's figure