Outlier Detection for Cognitive Diagnosis Models

Xinyue Zhao

Abstract

This paper introduces a method for outlier detection in cognitive diagnosis assessments. We apply the popular Cognitive Diagnosis Models (CDMs), in particular the DINA model. Through the paper, the process of outlier detection is demonstrated by using an Educational assessment dataset.

Keywords: Cognitive Diagnosis Models, Q-matrix, outlier detection.

1. Introduction

Cognitive Diagnosis Models (CDMs) are important statistical tools in psychometrics field, which provide conclusions about students' abilities. The Cognitive Diagnosis Model analytic process includes two steps: the qualitative step and the quantitative step. The first step of CDM analysis is qualitative: the test ability is divided into various basic abilities, termed as skills. After defining number and contents of skills, a Q-matrix is created and composed of skills required to master each item (Tatsuoka, 1983). Then, for the quantitative step, students are classified into dichotomous skill classes, which predict whether students present the defined skills. The results of CDMs are threefold: First of all, the distribution of skill classes is a combination of skills. Second, skill mastery probabilities indicate the proportion of students who master the individual skills. Finally, for each student, individual skill profile is constructed according to possession or nonpossession of skills (George, Robitzsch, Kiefer, Groß, and Ünlü, 2016). The CDMs have three general frames: the generalized-deterministic input noisy-and-gate (G-DINA; de la Torre, 2011) model, the generalized diagnostic model (GDM; von Davier, 2008) and the log-linear cognitive diagnosis model (LCDM; Henson, Templin, and Willse, 2009) (George, et al, 2016). The CDM framework associates various models which are mainly different in three aspects: compensability, dimensionality, and stochastic component (George, et al, 2016).

The statistical models are designed to simulate and analyze the performances of objects in reality. Even though these models are constructed for mimicking the real behaviors, data generated by the model can't be as same as the real situation. The points, where the responses are significantly different from the expected values from the model, are often called "outliers".

The outlier detection is essential in statistic data analysis. First of all, the number of outliers influences the goodness of fit of a designed model. The more outliers are detected in a dataset, the less fitted of the model. In other words, few outliers imply that the designed model is quite close to real data, representing the actual performances of observations well. Then, outliers can prove the validity of constructed models. The occurrence of outliers is often caused by either a flaw in statistical models or an experimental error. The modification of models can solve the flaw while the later problem can be resolved by removing those experiential inaccurate points from the original dataset and forming a new dataset. Finally, related with various subject fields backgrounds, outliers are important. The outliers in a dataset represent the abnormal behaviors in practice. For example, in the field of Education, outliers can represent the students' abnormal responses to designed questions. By learning and understanding them, we can improve and revise the method for reaching a better performance.

In cognitive diagnosis, the CDMs are widely used since they can incorporate with various models, which can be utilized in many fields. For example, the utilization of CDMs in field of Education helps experts understand the performances of students. The responses of students to different items assist experts to learn about the students' behaviors in a specific subject. The outliers in CDMs represent the abnormal behaviors of students. Experts can refine models and methods by interpreting the meanings of outliers.

In the following parts, more detailed contents related with CDMs are introduced. In Section 2, basic terminologies of CDM are demonstrated and outlier detection method is illustrated as well. In Section 3, simulation for outlier detection method is manifested. In Section 4, the application of outlier detection method is testified by one real dataset and a brief summary of the dataset is introduced. In Section 5, the interpretations of results and conclusions are summarized.

2. Model

2.1 Terminology

The dataset records students' performances in an achievement test. In the dataset, n students response to J items. A value of 1 means that students have a correct item response and a value of 0 means that students have a wrong item response. X_{ij} is a dichotomous manifest response of the ith student, i = 1, …, n, to the jth item, j = 0, …, J (George, et al, 2016). The ith of X matrix, X_i indicates the responses of the ith student to all J items and X_i is a response pattern of the ith student.

The K skills, which are needed for students to master each of J items under consideration, are defined by Education experts as α_k, k = 1, …, K (George, et al, 2016). In addition, a J× K dimension matrix is defined to record skills required to master items, namely, Q-matrix. In the Q-matrix, each element Q_{jk} has a value of 1 if the kth skill is related with mastery of the jth item. Otherwise, Q_{jk} has a value of 0. For items' condensation rules, they are determined and specified by experts.

Besides, each ith student's skill profile is denoted as $\boldsymbol{\alpha_i}$, i = 1, …, n. The student's skill profile describes each student's performance: he (she) presents or absents skills. Also, student's skill profiles form the basis of students' diagnostic feedbacks and provide bases for further instruction and learning (George, et al, 2016).

2.2 DINA

The DINA model is one of the most widely used and core of CDMs because of its simplicity and parsimony (George, et al, 2016). The DINA model has noncompensability rule: students have to possess all skills for an item in order to show successful mastery of the item (George, et al, 2016). The DINA model tests an individual student's probability to master an item with deterministic and probabilistic components.

For the deterministic component, the DINA model expresses the ith student's latent responses to the jth item dichotomously,

$$\eta_{ij} = \prod_{k=1}^{K} \alpha_{ik}^{q_{jk}},$$

where α_{ik} represents one element in the ith student's skill profile $\boldsymbol{\alpha_i} = [\alpha_{i1}, \dots, \alpha_{iK}]$ (George, et al, 2016). The value of η_{ij} equals to 1 indicates that a student possesses all or more than required skills is expected to master the item. Otherwise, the value of η_{ij}

equals to 0 represents that a student is not expected to master the item (George, et al, 2016).

For the probability component, the DINA model includes possible outcomes which are deviated from the expectations. A possible situation can be: the ith student possesses the skills and is expected to have the correct response to the jth item but he (she) may slip and fail the item. Alternatively, the ith student may have correct response to the jth item by a lucky guess even if he (she) does not process the skills required by the item. Then, let g_j and s_j denote the probabilities of guess and slip respectively. The numerical formulas for calculating probabilities are

$$g_j = P\left(X_{ij} = 1 \middle| \eta_{ij} = 0\right), j = 1, \dots, J$$

and

$$s_j = P\left(X_{ij} = 0 \middle| \eta_{ij} = 1\right), \; j = 1, \dots, J$$

To combine two probabilities together, the probability of the ith student succeed in the jth item is

$$P\left(X_{ij} = 1 \middle| \alpha_i, g_j, s_j\right) = (1 - s_j)^{\eta_{ij}} \cdot g_j^{(1-\eta_{ij})} = \begin{cases} g_j & if\ \eta_{ij} = 0 \\ 1 - s_j & if\ \eta_{ij} = 1 \end{cases}$$

Hence, from the formula shown above, the probability of the ith student succeed in the jth item is determined by two elements: the probability of guess g_j and the probability of non-slip $1 - s_j$. More explicitly, $\eta_{ij} = 0$ represents that all students who are not expected to master the item, and $\eta_{ij} = 1$ represents that all students who are expected to master the item. The probability of student' correct responses to the jth item is modeled by the chance that all non-mastered students guess the response correctly g_j and all mastered students solve the item without slipping (George, et al, 2016).

Except the DINA model demonstrated above, the CDMs include three other main models, the DINO model, the G-DINA model, and the GDM. The DINO model is similar to the DINA model but has a different principle in testing whether students are proficient in an item: the DINA model requires students to master all skills related with an item while the DINO model requests students to possess at least one skill included in an item. The G-DINA model is introduced by de la Torre to relax restrictive two-probability constraints (George, et al, 2016). The model shows the students' different probabilities of mastering items, which require various skills. Compared with the DINA and the DINO model, the G-DINA model is more complex, and it may be used to derive the DINA and the DINO model (George, et al, 2016). The GDM is a framework introduced by von Davier 2008 and contains most of CDMs for dichotomous and polytomous response data (George, et al, 2016).

2.3 Outlier detection method

The DINA model provides the regression for data. There are several model fit methods in CDMs to measure the goodness-of-fit. The commonest three methods are: the Akaike information criterion (AIC; Akaike, 1973, function AIC), the Bayesian information criterion (BIC; Schwarz, 1978, function BIC) and a χ^2 overall goodness-of-fit measure based on dependence between items (Chen and Thissen, 1997, function IRT.modelfit). The numerical formulas for AIC and BIC are

$$AIC = -2\log L(X) + 2\cdot p$$

and

$$BIC = -2\log L(X) + \log I \cdot p$$

where log L(X) is the maximal log-likelihood of the model and p is the total number of estimated parameters (George, et al, 2016). The criterion for AIC and BIC is that smaller the value, better the regression. For the χ^2 measure, it calculates the difference between probabilities of predicted values by models and observed values. Its criterion is that lower the value, better the fit.

For the comparison of different fitted models, an ANOVA tests can be used. In addition, the correlation of two variables is one of influential factors for regression model. The correlation is mathematically defined as

$\mathrm{Cor}(x,y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$ (James, Witten, Hastie, and Tibshirani, 2017),

where $y_i - \bar{y}$ is the difference between the response value and the predicted value. The difference is called “residual”. From the residual plot, we can observe the patterns of plots and figure out the obvious outliers directly.

The process of outlier detection method is illustrated below. First of all, we fit the data from a specific dataset with the DINA model. Then, from the summary of the DINA model, the values of AIC, BIC and the values of guess, slip and proportion can be obtained. In order to detect the outliers, we remove a data point from the dataset each time and refit the DINA model. Then, new values of guess, slip and proportion are calculated. Both the differences and the norms of differences of guess, slip and proportion are computed. The norm measures the distance between data's observed values and predicted values by model. The numerical formula of norm is

$$\parallel \beta \parallel_2 = \sqrt{\sum \beta_i^2} \ ,$$

where β is a set of vectors (James, et al, 2017). With a set of differences of guess, slip and proportion, plots are made by setting x-axis "student id number" and y-axis "norm of difference". From the plots, the potential outliers, which display far distances from other points, can be recognized. Since items are dependence of others, the values of guess, slip, and proportion have correlations. In other words, the point that has large difference in guess may also have large difference in slip or proportion. Thus, three correlation plots are made.

The results gained from plots may not be convincing. The responses of each potential outlier point are extracted from the dataset. By understanding their performances to items, the potential outliers can be determined whether they are outliers or not.

However, some improvements of method are applied so that the results of detecting outliers are more convincing and more precise. The individual t-tests of each point are conducted. The t-test is a hypothesis test to determine whether a dataset follows the t-distribution under the null hypothesis. The null hypothesis is an assumption which we suppose before do the hypothesis test. Also, a significant level is the probability that the null hypothesis is rejected given it is true. A rule should be specified: the smaller significant level is, the more precise result is. The t-value of a result, t, is the probability that the result obtained at least extreme, given the null hypothesis is true. The formula for t-value is

$$\mathrm{t} = \frac{\bar{\mathrm{x}} - \mu}{\mathrm{s}} \ ,$$

where μ, s represent the mean and standard deviation for the dataset. In this paper, the null hypothesis for the t-test is constructed that "The point is not an outlier". The alternative hypothesis is "The point is an outlier". The significant level is $\alpha = 0.05$. If t-value of a point is larger than the significant level α, the null hypothesis is rejected, which implies that the point is an outlier. Otherwise, the point is not an outlier.

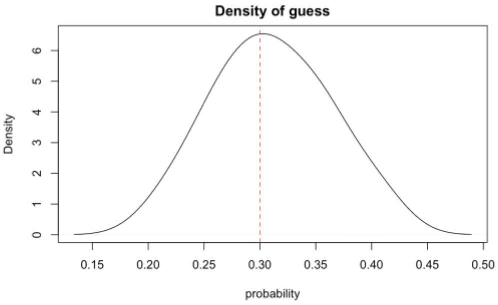
Since multiple t-tests are utilized in this paper, the Bonferroni correction is needed. The Bonferroni Correction adjusts each individual hypothesis test by increasing the significant level as $\alpha^* = \frac{\alpha}{n}$. With new significant level α^*, results of individual t-tests are reported.

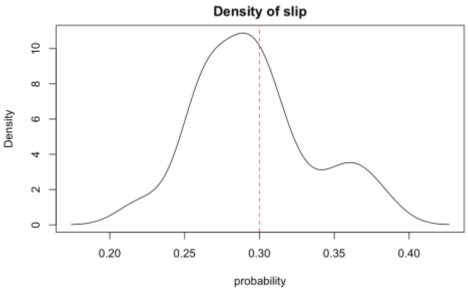
3.Simulation

Simulation is one of core processes of statistical analysis. Compared with method application, simulation is designed for proving the validity of method. For simulation, control is a key component related to the fidelity of a non-random process to a random process (Gentle, 2009). More intuitively, the data are generated on purpose rather than obtained from the real experiments. With the assumed dataset, the outlier detection method is applied and observed whether the results are same as the expectation. The expectation of simulation is that all outliers generated in the dataset can be detected through the outlier detection process designed and formulated previously. Since all data in the dataset are made up, the outlier points have already been known before the simulation. Then, we compare the outcomes obtained from the simulation with the prediction. If they are matched, the method used for detecting outliers is convincing.

In CDMs, the dataset for the DINA model can be simulated by an existing function in the CDM R- package. The data simulation tool for the DINA model, *sim.din,* is utilized to simulate dichotomous response data based on a CDM (Package 'CDM', 2018). In data simulation, the number of observations, the Q-matrix, the values of guess and slip, the mean, the sigma and the rule are elements required to be specified for using formula. Also, the value of alpha, the matrix of attribute patterns given as an input rather than underlying latent variables (Package 'CDM', 2018), is selected from NULL or non-Null. The mean and the sigma would be ignored by setting the alpha non-NULL.

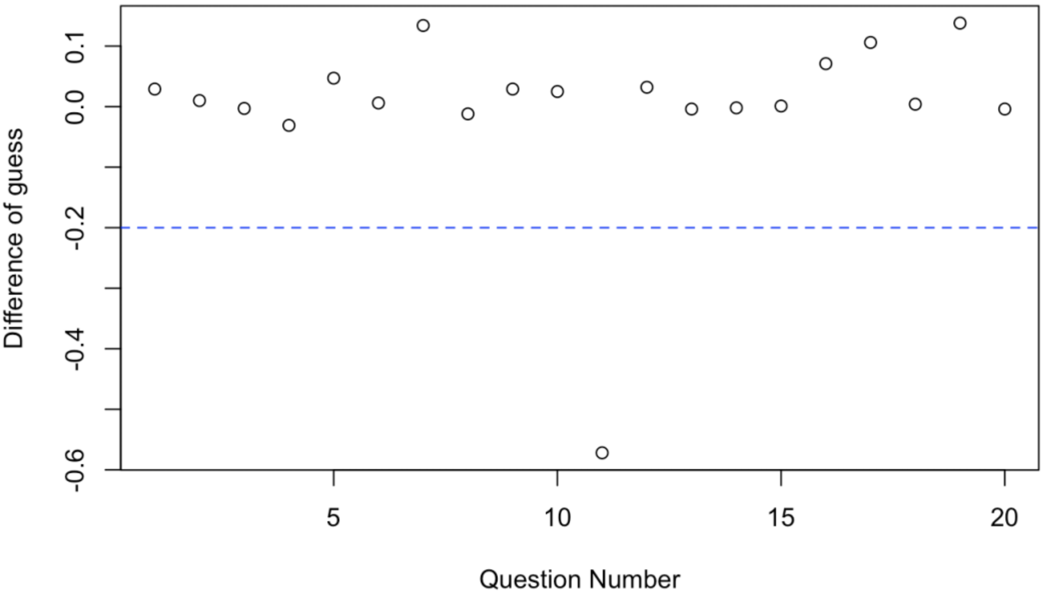
The data simulated for proving validity of outlier detection method is assumed that 500 observations and 20 items are in the dataset since the real dataset includes 536 students and 11 skills. The values of guess and slip are determined both 0.3, which imply that only 30% responses of students are either slipped or guessed. Then, one of important elements in the function is the mean. The default value is 0. In order to guarantee the randomness of dataset, the value of mean is generated by Bernoulli trials. The Bernoulli trial is the special case of the Binomial distribution. The Binomial distribution consists n independent Bernoulli trails. A random generator for Binomial distribution is included in the R-package. Then, a random Bernoulli trial can be operated by setting n = 1.





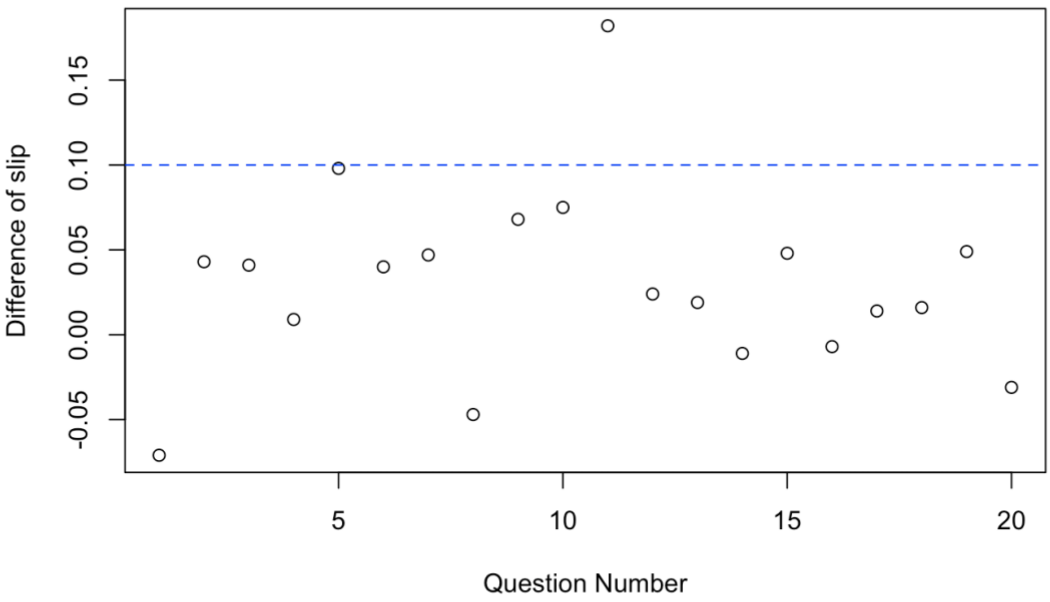
The x-axis represents the probability of values of guess (slip). The y-axis represents the density of values of guess (slip). From two plots shown above, both curves are symmetric about probability = 0.3. Since the curves of Binomial distribution are symmetric, and are bell-shaped about 0, the generated plots are conformed with our expectation.

First, a specific point in the dataset is chosen to be an outlier. From the method mentioned from previous part, the same process is conducted here. The 11th point in the dataset is selected as an outlier, by setting the probability = 0.9.



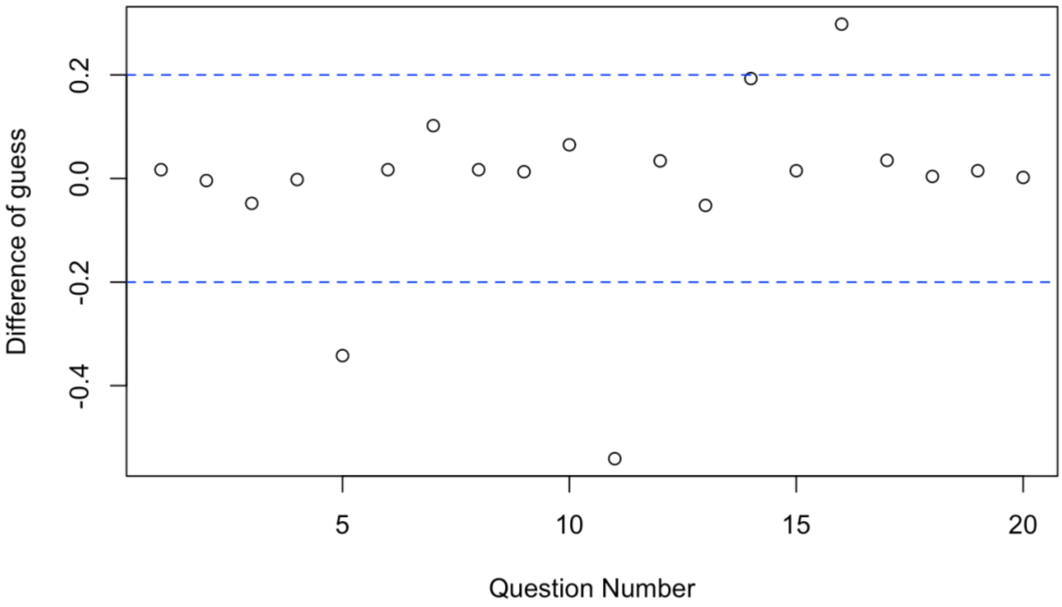
The plot above shows the differences of each questions' guess values. The x-axis represents the number of questions, from 1 to 20. The y-axis is the differences of values of guess, equivalently, the differences between the original guess values which are expected from the model and the new guess values from the new Q-matrix. From the plot, the difference of 11th point is far from other points. Hence, the 11th point can be detected as an outlier. Moreover, the method for outlier detection is valid.

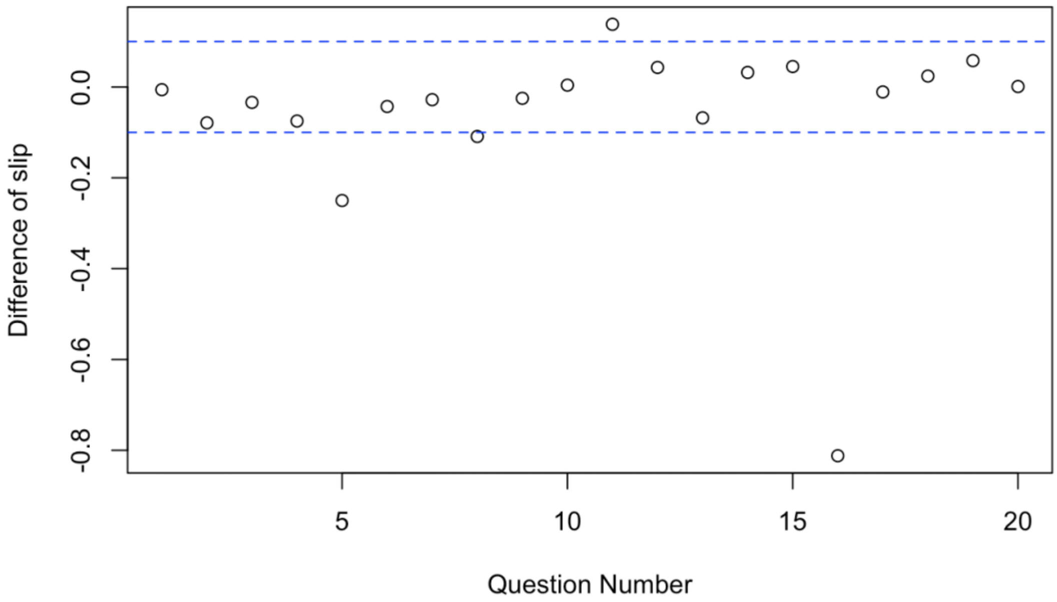
Same process is repeated for parameter slip and proportion.



Then, the x-axis for this plot still represents the question number and the y-axis is the differences of values of slip. From the dashed line shown above, differences of all other points fall within 0.1, except the 11th point. The 11th point shows an extreme distance from other points. Thus, the 11th point is an outlier, same as our assumption.

In addition, more outliers are added in the dataset to prove the validity of method for detection. The 5th, the 11th and the 16th point are selected as outliers by assuming the probability = 0.6, 0.9 and 0.1 respectively. The outlier detection is repeated again.





From two plots shown above, the x-axis represents the question numbers, the y-axis is differences of guess and slip respectively. From the plots shown above, three points in the dataset have large differences of guess and slip compared with other points. These three points, the 5th, the 11th, and the 16th point, can be determined as outliers. The result is consistent with our assumption. Thus, the method for outlier detection works.

4. Outlier Detection

4.1 Dataset

In this paper, one of datasets is studied and is detected with outlier detection method. The dataset is the second of five datasets in "fraction dataset". The responses of 536 students to 11 items are included in the dataset (Package 'CDM', 2018). The dataset contains three Q-matrices which relate to different skills. The first Q-matrix is the Q-matrix of Henson, Templin and Willse (2009), and the second Q-matrix is the alternative one in the de la Torre (2009) (Package 'CDM', 2018).

	H01	H02	H03	H04	H05	H06	H08	H09	H10	H11	H13
1	1	1	0	1	1	1	1	1	0	1	1

2	1	1	1	1	1	1	1	1	1	1	1
3	1	1	0	0	0	0	1	1	0	0	0
4	1	0	0	1	1	1	0	1	0	1	1
5	1	0	0	0	0	1	0	0	0	0	0
6	0	0	1	0	0	0	0	0	0	0	0
7	0	0	1	0	0	0	0	0	0	0	0
8	1	0	0	0	0	0	0	0	0	0	0
9	0	0	1	0	0	1	0	0	0	0	0
10	0	0	1	0	0	1	0	0	0	0	0

The table shown above is first ten students' responses to eleven questions. The horizontal row is each student's response to all eleven items. The vertical lines are eleven items. The eleven items in the dataset are "H01", "H02", "H03", "H04", "H05", "H06", "H08", "H09", "H10", "H11" and "H13". For example, the first student has all correct responses to items, except item H03, and H10.

The detection is divided into three cases according to three different Q-matrices. Every Q-matrix shows the relationship of different skills. The Q-matrix1 shows the relationship between eleven items and three skills "QH1", "QH2" and "QH3". The second Q-matrix shows the relationship between eleven items and five skills "QT1", "QT2", "QT3", "QT4" and "QT5". Similarly, the Q-matrix3 reflects the relationship between eleven items and three skills "Dim1", "Dim2", and "Dim3". The first Q-matrix is shown as an example.

	QH1	QH2	QH3
H01	1	1	0
H02	1	0	1
H03	1	0	1
H04	1	0	0
H05	1	1	0
H06	1	1	0
H08	1	0	1
H09	1	0	1

H10	1	0	0
H11	1	0	0
H13	1	1	0

The row is the relationship between each item with three skills. The columns are three skills “QH1”, “QH2” and “QH3”. For example, the correct response to item “H01” is expected to master the skills “QH1” and “QH2”. The interpretations of three skills are “borrowing from a whole number”, “separating a whole number from a fraction” and “determining a common denominator” (Henson, Templin and Willse, 2009).

First of all, we fit the data with the DINA model. The summary of the DINA test is given below.

```
Deviance =    5475.249   |     Log-Likelihood = -2737.624


Number of item parameters: 22
Number of skill class parameters: 7


Information criteria:
   AIC =    5533.249
   BIC =    5657.489


Mean of RMSEA item fit: 0.054
```

Since for the RMSEA, when its value approaches to 0, the model fits worst; when its value approaches to 1, the model fits best. Hence, the value of RMSEA is 0.054, the DINA model is not fitted well. Thus, there are outliers in the dataset.

```
Marginal skill probabilities:
           skill.prob
QH1          0.4893
QH2          0.7258
QH3          0.6660
```

From the table shown above, the probabilities of mastering each skill is demonstrated. For example, the value of skill probability of “QH1” is 0.4893, which implies that 48.93% of students can master the skill “borrowing from a whole number”.

```
Tetrachoric correlations among skill dimensions
          QH1       QH2      QH3
QH1   1.0000    0.8148   0.5596
QH2   0.8148    1.0000   0.2853
QH3   0.5596    0.2853   1.0000
```

This table provides us correlations between three skills. Among all three skills, the skill “QH1” is highly related with the skill “QH2” and the skill “QH2” is less related with the skill “QH3”. For the interpretation, the skill “borrowing from a whole number” is highly related with the skill “separating a whole number from a fraction”. The skill “separating a whole number from a fraction” is less related with the skill “determining a common denominator”.

```
Skill Pattern Probabilities
     000       100       010       001       110       101       011       111
 0.12767   0.00009   0.12767   0.12767   0.07857   0.01876   0.12767   0.39190
```

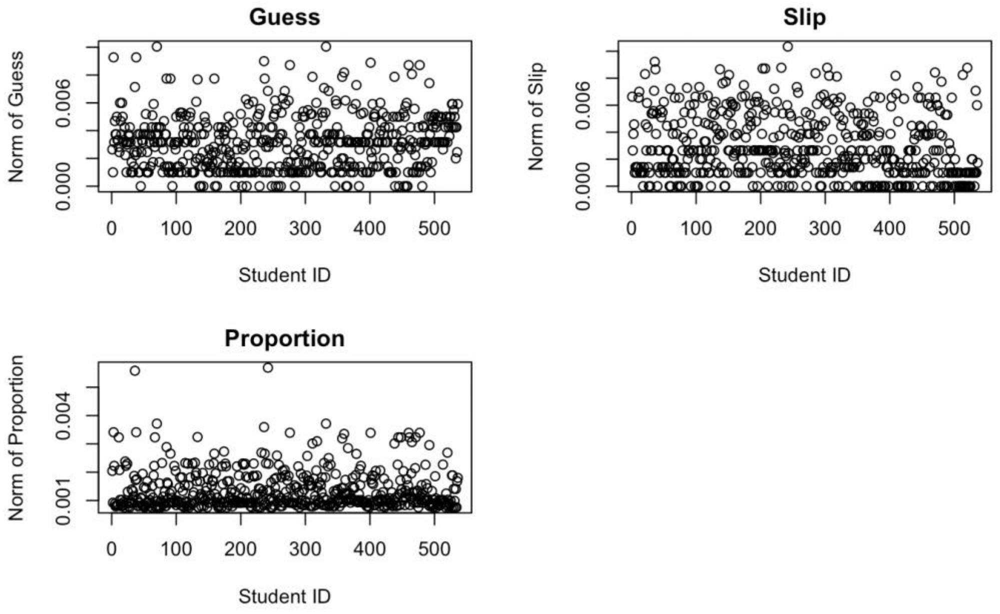
The table above shows the proportions of students’ skill patterns. For example, 12.767% students fail all three skills.

Similar to the information obtained above, among all skills, the skill “QT3” is the commonest skill which students master while the skill “Dim3” is the hardest one for students to master. Also, for correlations, the skill “Dim1” and “Dim2” are highest correlated while the skill “QT1” and the skill “QT4” are least correlated.

4.2 Application of method to a dataset fraction2

4.2.1 Outlier Detection

Follow the process illustrated in Section 2.3, the plots of Case I are shown below.

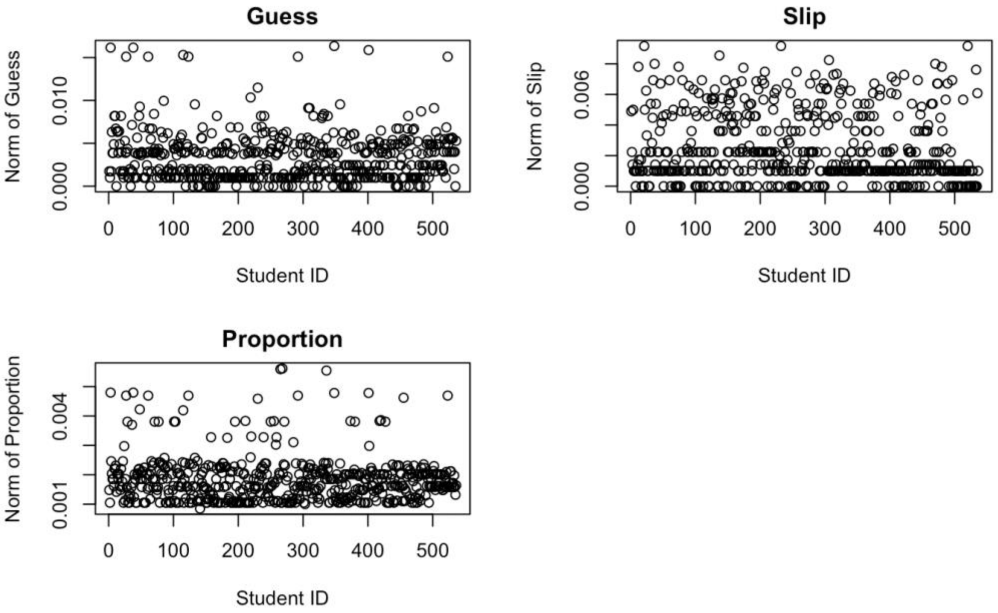


The x-axis of three plots is “student’s ID number” and the y-axis is “norms of differences of guess, slip and proportion” respectively. Based on the plots, we can observe the patterns of points and concludes some potential outliers.

The plot of guess doesn’t show clear potential outliers. There are two obvious potential outliers for the parameter proportion and an obvious potential outlier for the parameter slip. The potential outlier for the parameter slip is the 242nd point. The 36th and the 242nd point are two potential outliers for the parameter proportion.

To summarize: There are two potential outliers, the 36th and the 242nd point, in the dataset from the first Q-matrix.

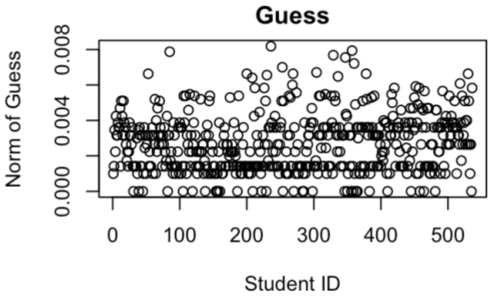
Case II:

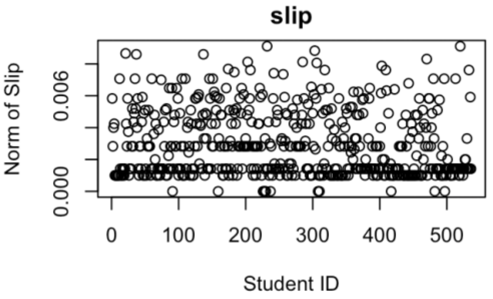


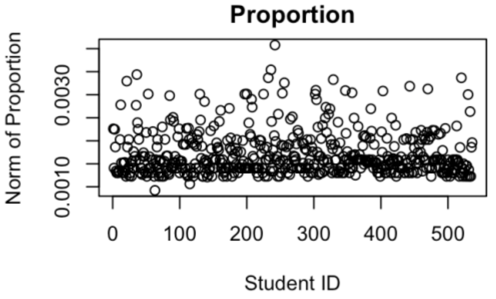
Similar to Case I, the x-axis and y-axis are "student's ID numbers" and "norms of differences of guess, slip and proportion" respectively. From three plots, potential outliers can be concluded. There are ten potential outliers for the parameter guess: the 3rd, the 27th, the 38th, the 61st, the 115th, the 123rd, the 292nd, the 348th, the 401st, the 523rd point. There are three obvious outliers for the parameter proportion, the 265th, the 268th, and the 336th point. However, no obvious outliers can be observed from the plot for the parameter slip.

To summarize: Thirteen potential outliers can be detected from plots in the second Q-matrix. The 3rd, the 27th, the 38th, the 61st, the 115th, the 123rd, the 265th, the 268th, the 292nd, the 336th, the 348th, the 401st, and the 523rd point are potential outliers for Case II.

Case III







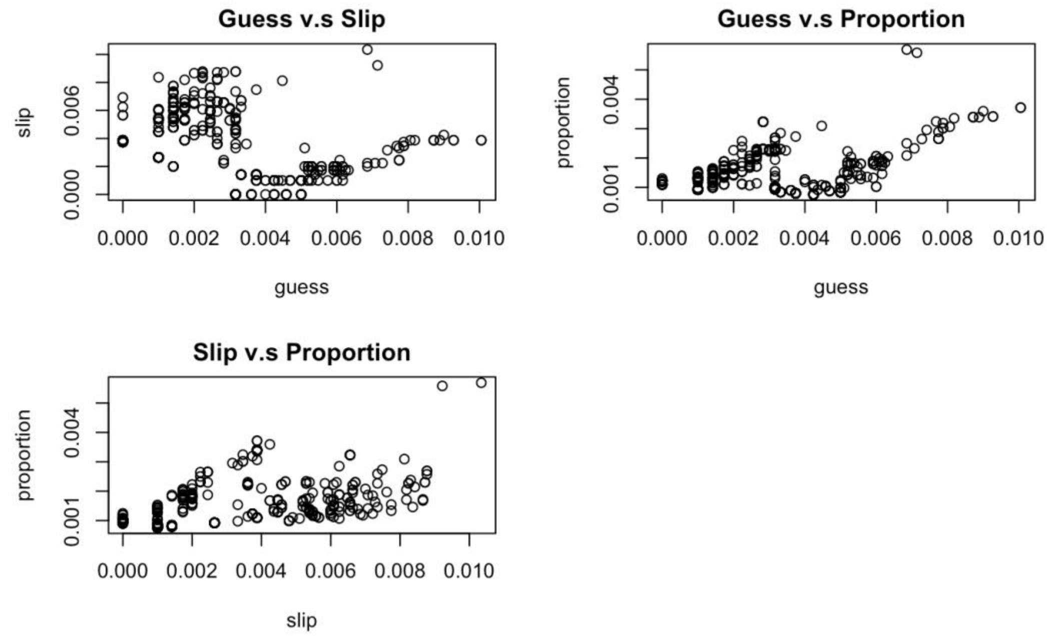
The plots obtained in Case III are similar to the plots shown in Case I and II. “Students’ ID numbers” is x-axis and “norms of differences of guess, slip and proportion” is y-axis accordingly. The points in plots of the parameter guess and the parameter slip are scattered. Hence, no obvious outliers can be detected. There is a potential outlier in plot of the parameter proportion, the 242nd point.

To summarize: An obvious potential outlier, the 242nd point, is observed from three plots in Case III.

From the results shown above, a further detection is conducted in order to test whether the point is a real outlier. Since there are overlaps in three cases, the potential outlier for the parameter guess may also be the potential outlier for the parameter slip and proportion. Thus, every two variables are plotted.

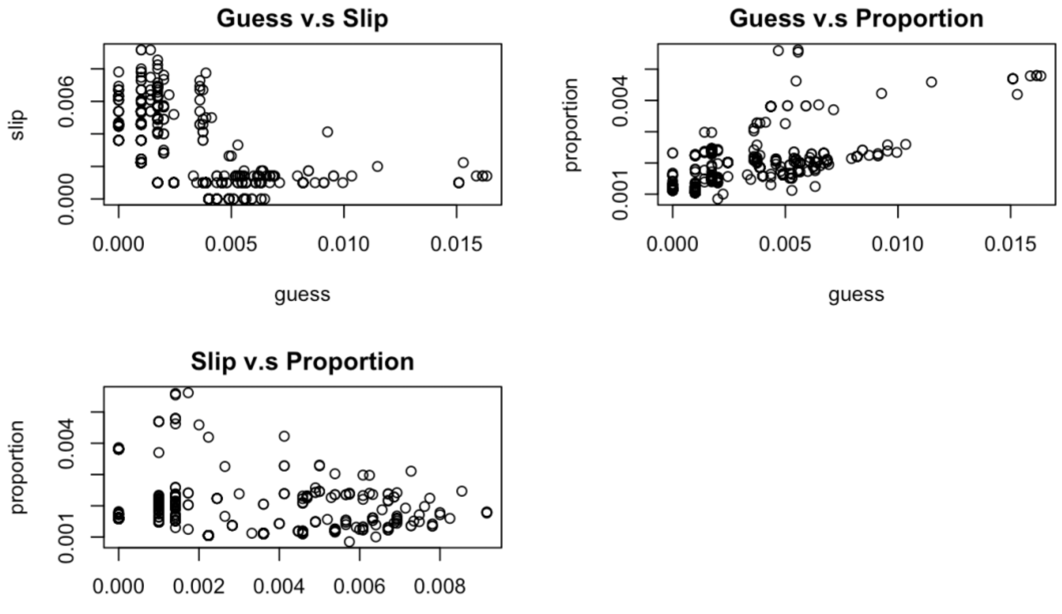
Three plots are completed for comparing. In the first plot, the values of the parameter guess and the parameter slip are compared. The x-axis is “the value of guess” and the y-axis is “the value of slip”. In the second plot, “the value of guess” is on x-axis and “the value of proportion” is on y-axis. For the last plot, the comparison between the parameter slip and the parameter proportion is shown by setting “the value of slip” as the x-axis.

Case I:



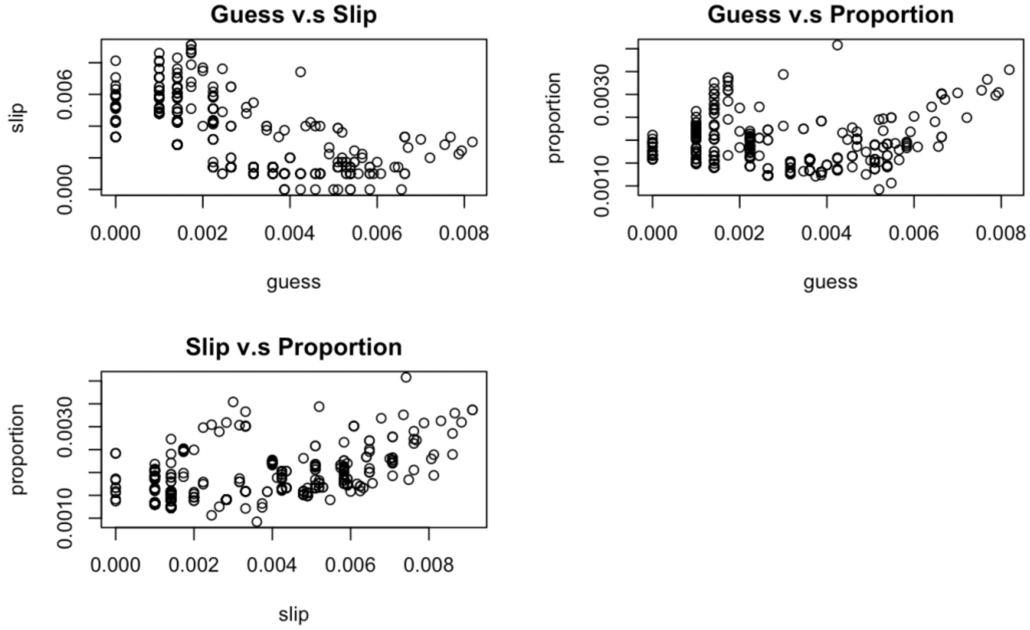
In the first plot, two obvious potential outliers are displayed on the right top corner, which imply that two points in the dataset have large values of both guess and slip. Also, two points with large values of both proportion and guess can be recognized as potential outliers in the second plot. For the third plot, two obvious potential outliers present abnormal performances, with large values of both slip and proportion.

Case II:



For the data in second Q-matrix, there is an obvious outlier cluster in the first and second plots. There are some potential outliers with large values of guess, slip and proportion in the dataset. However, from the third plot, no point has large values of both slip and proportion.

Case III:



For the third Q-matrix, the patterns of values of guess and slip are normal. There are some points have large values of both guess and proportion in the second plot. Also, a point with comparative large value of slip and proportion may be considered as a potential outlier.

From the plots, some points have large values in two variables. Hence, these points may be reasonable to be detected as outliers. However, the plots provide only a general view of potential outliers. Every performance of potential outliers should be analyzed to find the real outliers. The number of correct responses to items can provide an evidence for whether the point should be treated as an outlier.

The 21st point from the second Q-matrix is shown as an example.

	H01	H02	H03	H04	H05	H06	H08	H09	H10	H11	H13
21	1	1	1	0	1	1	0	1	1	1	0

From the table above, the 21st point succeeds eight items. However, majority of points have correct responses to four items.

By analyzing each potential outliers' responses, a summary is concluded. For Case I, there are two outliers, the 36th and the 242nd point. For Case II, there are eight outliers, the 21st, the 27th, the 61st, the 123rd, the 232nd, the 292nd, the 520th and the 523rd point should be real outliers. For Case III, only an outlier, the 242nd point. All points mentioned have either too many or too less correct responses to items.

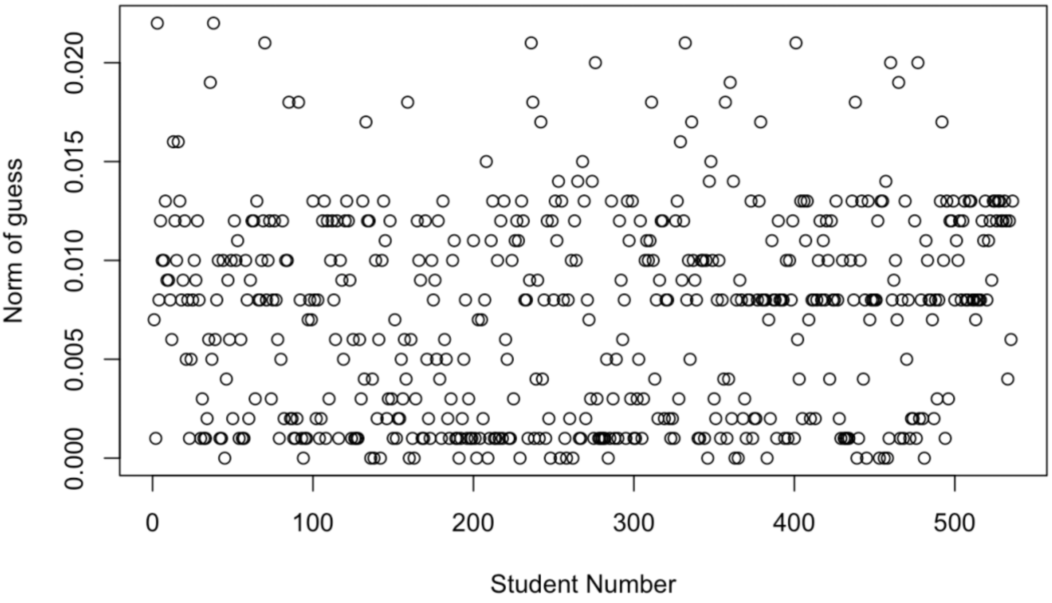
4.2.2 The individual point's T-test

Since compared with concluding the outliers by deriving from plots directly, detecting outliers with numerical results is more convincing. Each point in the dataset does the individual t-test. By limiting various significant levels, the number of outliers can be detected. In each Q-matrix, three parameters, guess, slip and proportion, are tested individually.

Case I:

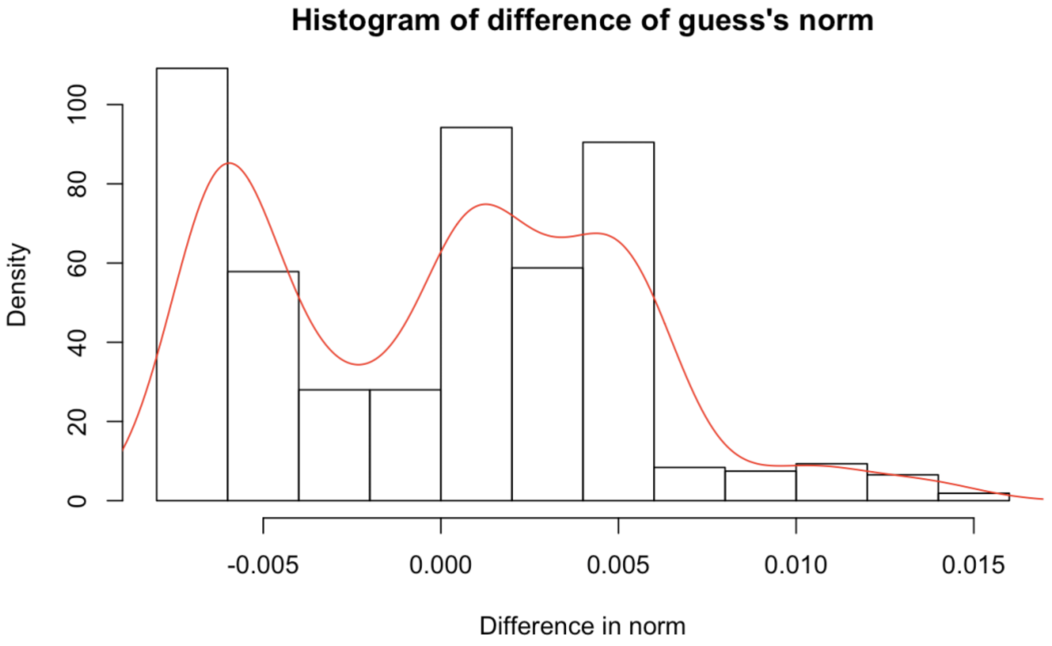
(i) Guess

First of all, we abstract the values of parameter guess from the whole dataset and store them in a vector. Then, we calculate the norm the parameter guess. We plot the points.



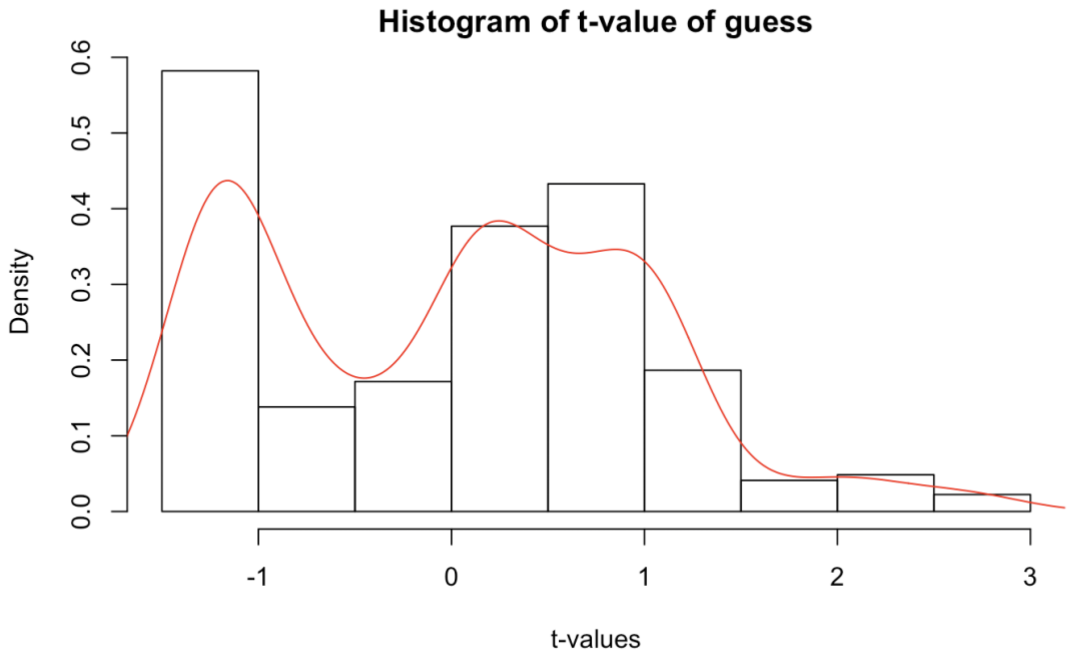
The x-axis is "student ID number", from 1 to 536. The values of norm of guess are made of the y-axis. From the plot, the points are scattered and show no pattern. Thus, concluding which point is a potential outlier directly from the plot is hard.

Secondly, we calculate the mean, the variance and the standard deviation. Then, we calculate the differences and plot the differences of every point and observe the pattern. In order to observe clearly, the density function is also plotted.

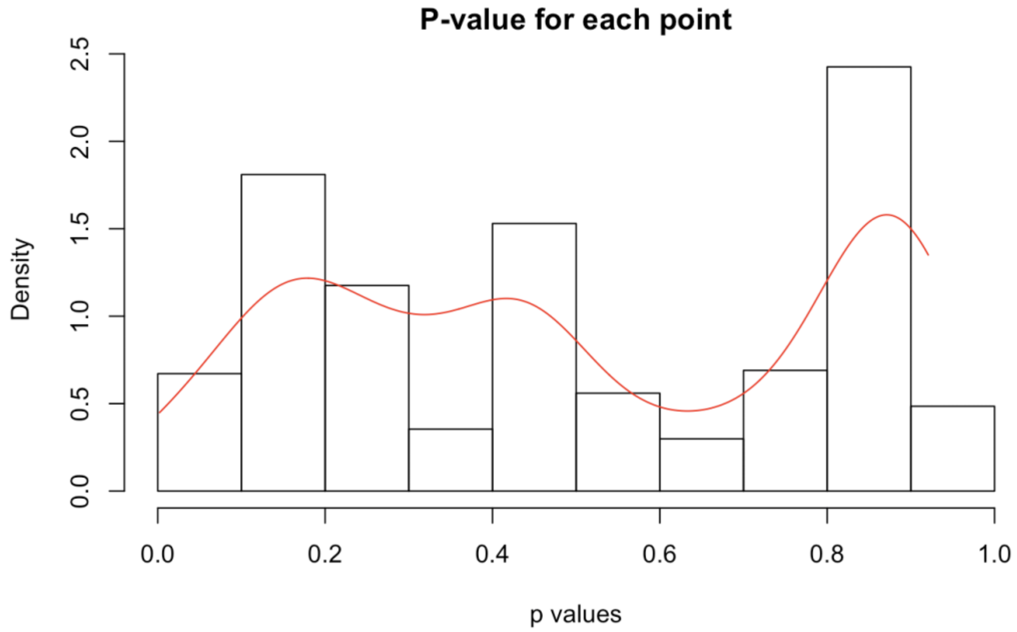


The x-axis of histogram is the difference in norm of guess. The y-axis is density. If there is no outlier in parameter guess, for the first Q-matrix, the density curve would be a bell-shaped, symmetric curve. Just as what should be expected, the pattern of plot is left-skewed and the density curve does not follow the Normal Distribution (0,1). Hence, for the parameter guess, there may be some outliers.

Then, t-values of each point are calculated. For observing easily, the histogram of t-values and the density curve are both plotted.



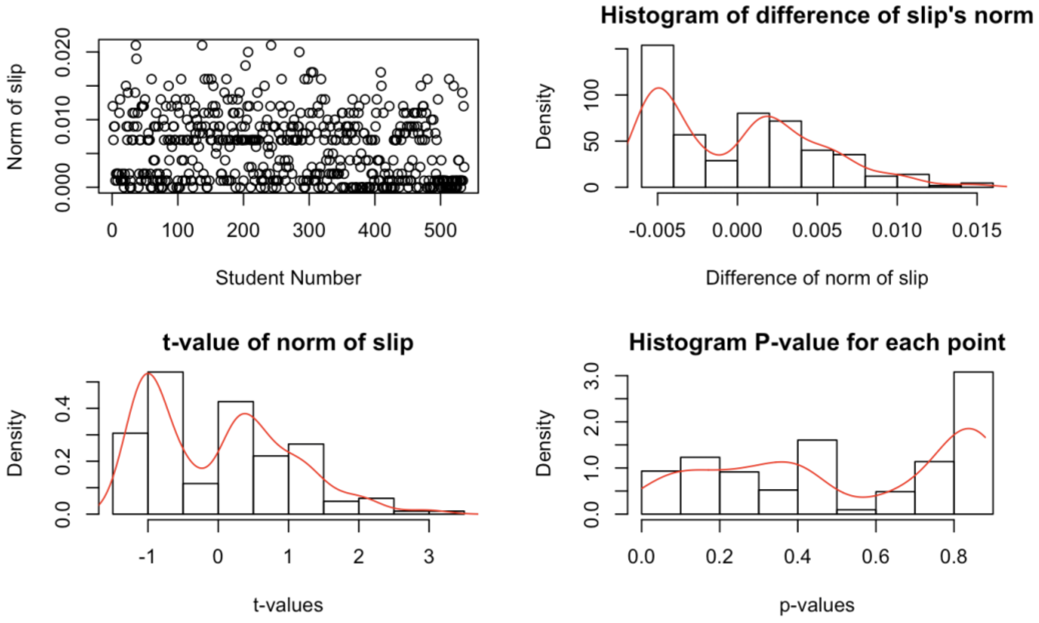
Also, with t-values, we can find every point's corresponding p-values. We obtain the p-values for every point. Then, we plot the histograms for them. In order to show the pattern clearly, the density function is used.



Because only outliers are cared about, all p-values points which are smaller than the significant level 0.05 are selected. We want to select the outliers, which imply to reject the null hypothesis. Since when the p-values are smaller than the significant level α, we reject the null hypothesis. With such criteria, twenty-six points are detected as potential outliers. By setting the significant level to 0.01, nine points are detected as potential outliers.

(ii) "Slip"

The process to test the parameter slip is quite similar to the one of the parameter guess.

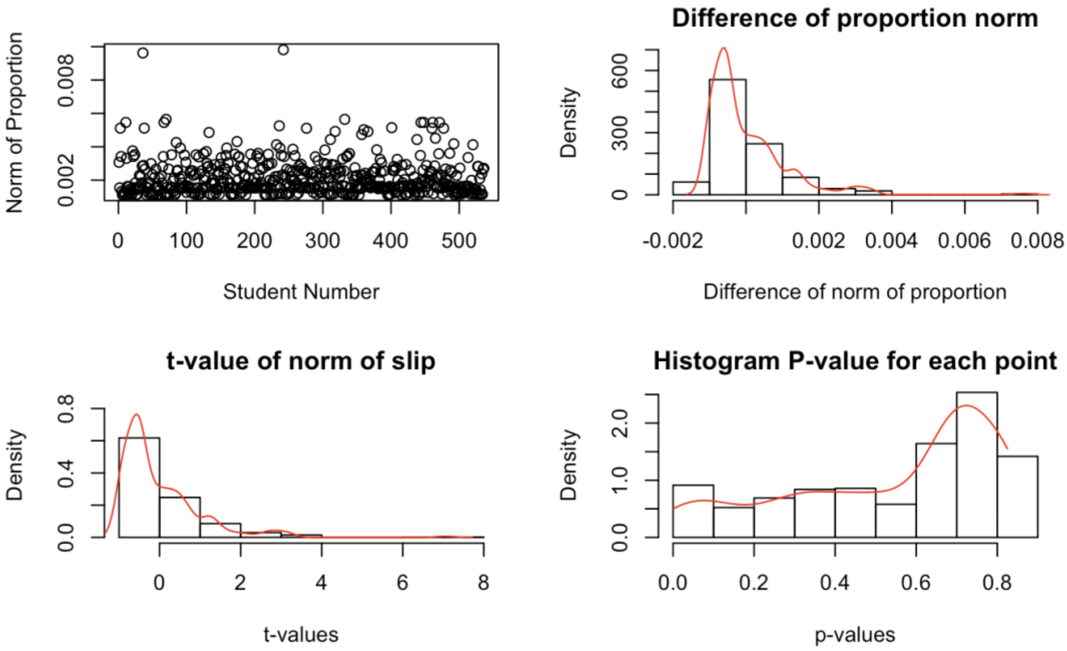


The first of four plots is the scatter plot of norm of the parameter slip. The x-axis is "student ID number" and the y-axis is "the norm of slip". From the plot, we can find that points are scattered and there exist some outliers.

The other three plots are histograms of differences, t-values of norm of slip and each point's p-value. The x-axis is "the differences of norm of slip, t-values, and p-values" respectively. The y-axis is density. All density curves in three plots show abnormal patterns. The curves of differences of norm of slip and t-values are left-skewed. Thus, outliers exist in the data of the parameter slip in first Matrix.

By setting the significant level $\alpha = 0.05$, the outliers are the points whose p-values are smaller than 0.05, failing to reject the null hypothesis. There are thirty-five potential outliers. If the significant level decreases to $\alpha = 0.01$, seven points are detected as potential outlier points.

(iii) Proportion



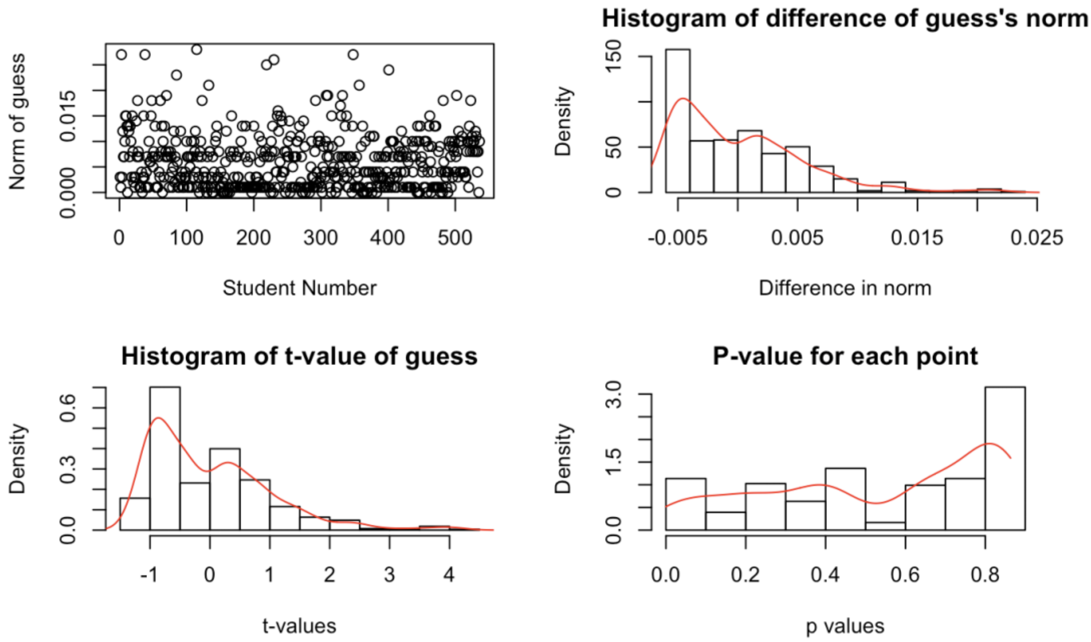
Similar to the parameter guess and the parameter slip, the same process is repeated for testing the parameter proportion. The first plot is the scatter plot of student's norms of proportion. In the plot, two obvious outliers are shown. With the significant level $\alpha = 0.05$, thirty-three points are detected as potential outlier points. If the criterion narrows to 0.01, twenty points would be considered as potential outliers.

Furthermore, for the first matrix, nine points are potential outliers for both the parameter guess and the parameter proportion. Seven points are potential outliers for the parameter slip.

Since all the detection processes for both the second Q-matrix and the third Q-matrix are same as the one for the first Q-matrix, which is demonstrated above, only the results and plots are shown in II, III cases. The first of four plots is the scatter plot of norms of guess, slip or proportion. The x-axis is "student ID number" and the y-axis is "the norm of guess, slip, or proportion". The x-axis is "difference of norm of guess, slip or proportion, t-values, and p-values" respectively and the y-axis is density for rest three plots.

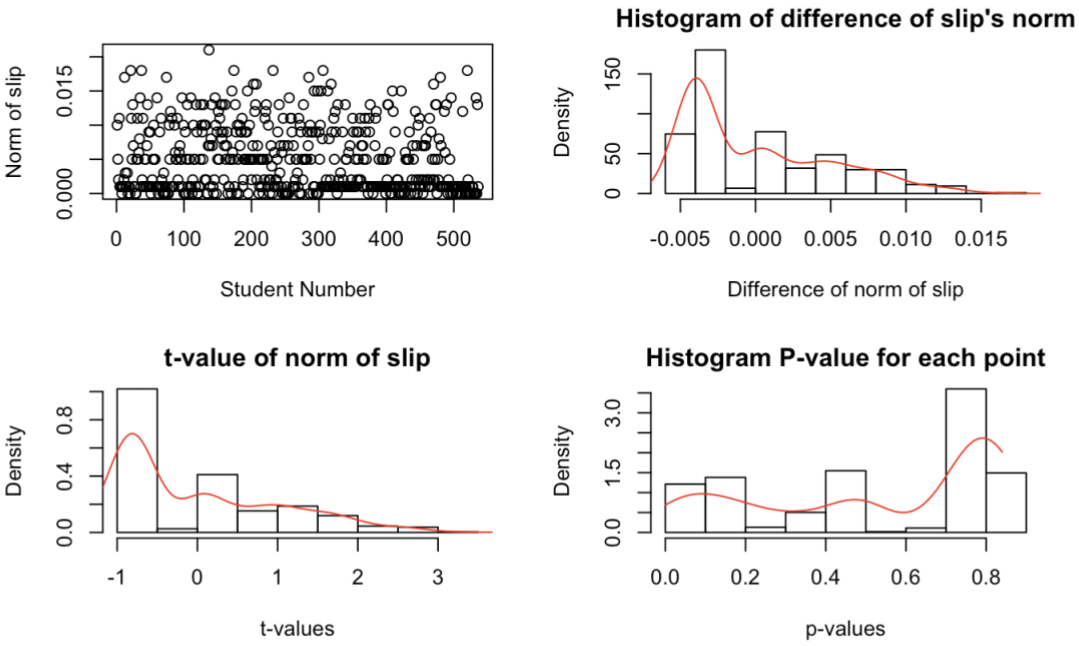
Case II

(i) Guess



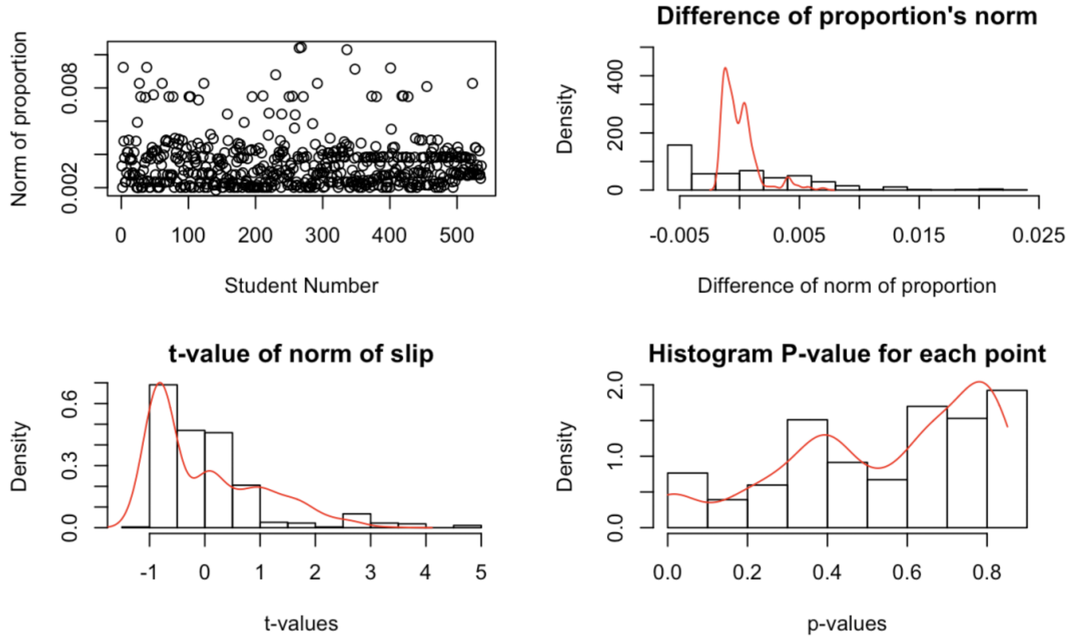
Both the scatter plot and density curves imply the existence of outliers. After a series of calculations, with the 0.05 as the significant level, thirty-six points are potential outliers. With 0.01 significant level, fifteen potential outliers are detected.

(ii) Slip



From the plots shown above, the density curves are left-skewed, which reflect the existence of outliers. With the 0.05 as significant level as criterion, fifty-five potential points are outliers. If the criterion narrows to 0.01, the number of potential outlier points is eleven.

(iii) Proportion

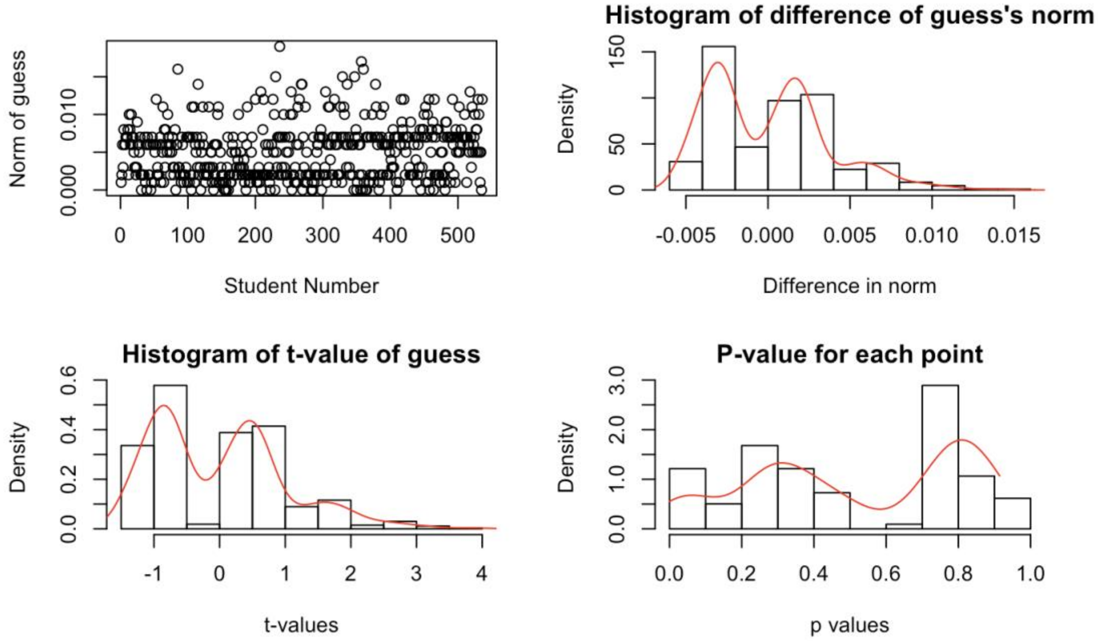


The first plot is the scatter plot of student's norms of proportion. There are several points, which have large differences from other points in the plots. The density curves in the second and the third plot are left-skewed, which also imply that some outliers are in the dataset. With 0.05 significant level, thirty-six points are potential outliers. Just as the process done above, changing the criterion to 0.01, thirty-two potential outliers are detected in dataset.

Specifically, five points are detected as potential outliers for both the parameter guess and the parameter slip.

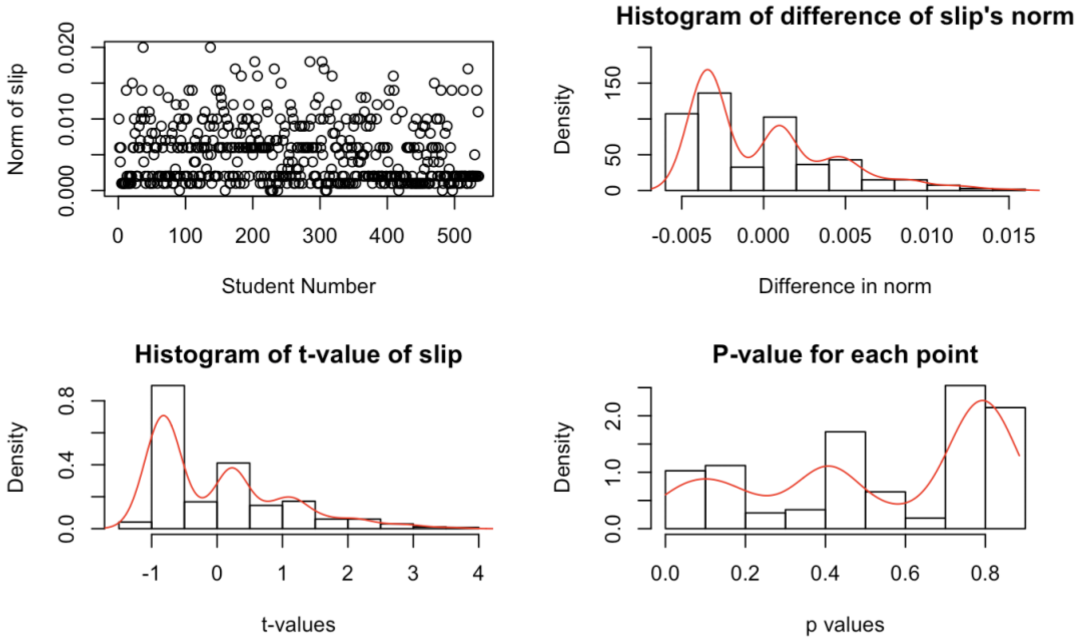
Case III

(i) Guess



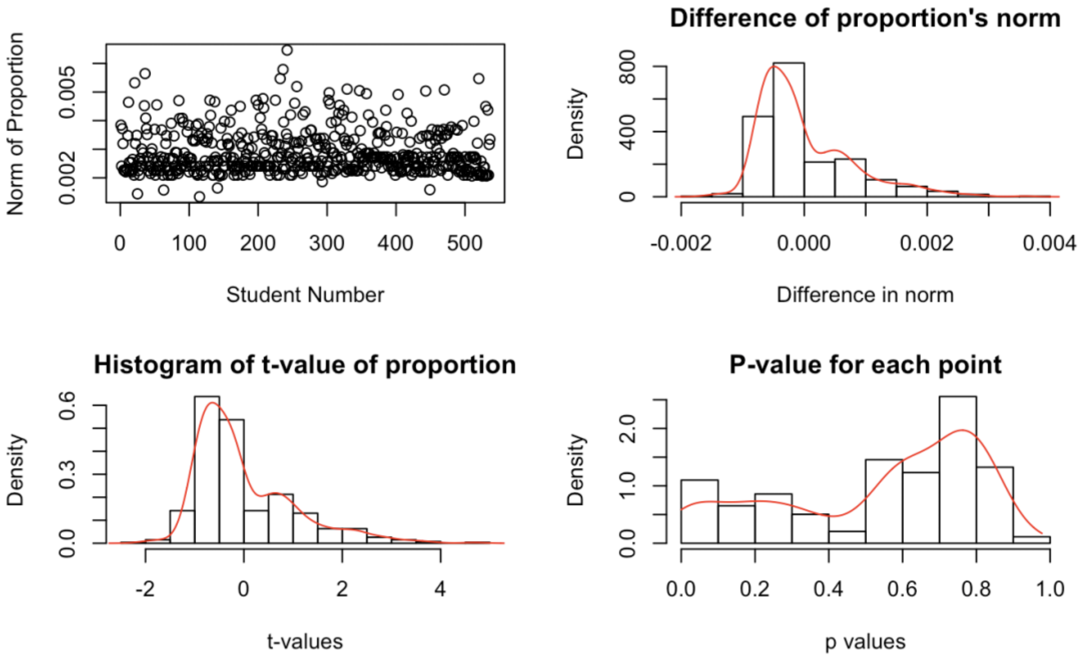
For the parameter guess, with the 0.05 significant level, forty-four points are detected as potential outliers. By changing the criterion to 0.01, twelve points are potential outliers.

(ii) Slip



For the parameter slip, with the 0.05 significant level, forty-five points are detected as potential outliers. If 0.01 is used as the criterion, the number of outliers decreases: sixteen points are tested as potential outlier points.

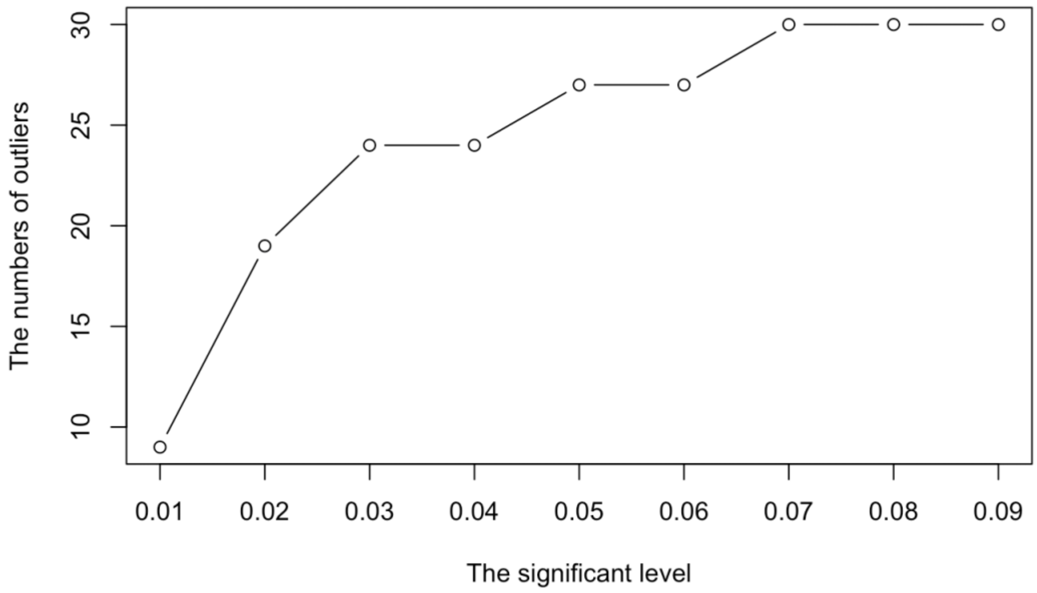
(iii) Proportion



In the first plot, several points show large differences from other points. For the parameter proportion, with the 0.05 significant level, forty-three points are detected are potential outliers. The number of outliers is less when the significant level is 0.01. In such condition, twenty points are potential outliers.

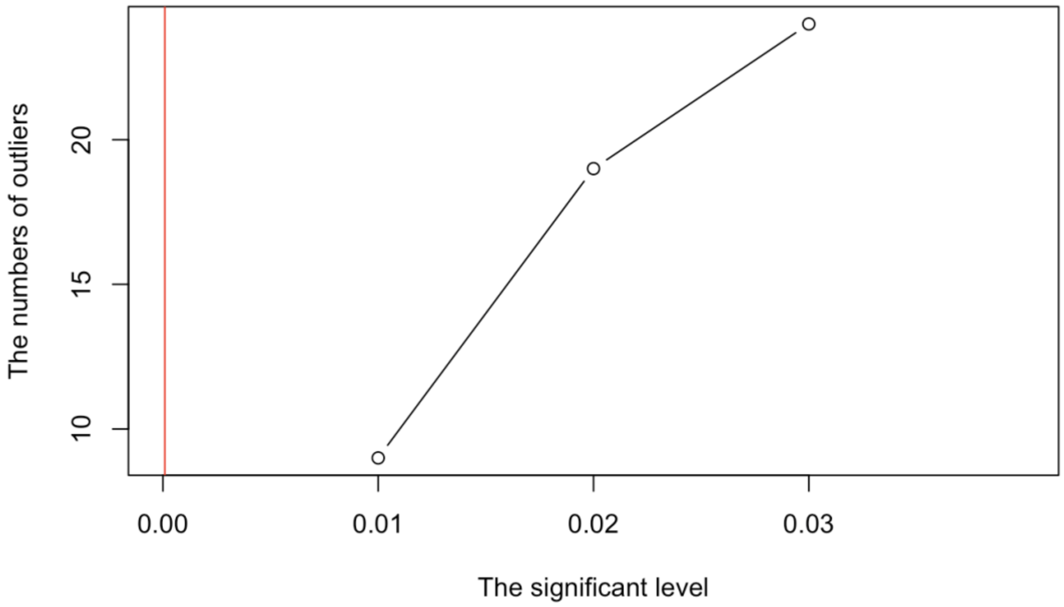
Similar to Case I and II, some points are repeatedly counting as potential outliers for two parameters. Five points are potential outliers for both the parameter guess and the parameter slip. Also, seven points are potential outliers for both the parameter slip and the parameter proportion.

In addition, the change of the number of outliers related with decreasing significant levels is also detected. The parameter guess of the first matrix is used as an example to be tested.



The x-axis of plot is the significant level. The y-axis of plot is the number of outliers. From the plot, the number of outliers approximately shows the increasing trend as the criterion becomes larger.

Furthermore, the Bonferroni Correction is also applied. The number of samples is 536. Hence, the new significant level is $\alpha^* = \frac{\alpha}{n}$, which is 0.00009. Unfortunately, no point falls in such small criteria for the parameter guess in the first matrix.



The significant level is labeled on the x-axis. Since the significant level of Bonferroni Correction is too small, only the interval [0.01, 0.03] is shown on the x-axis. The y-axis is the number of outliers. The red line represents the value of Bonferroni Correction's significant level.

4.2.3 Bonferroni Correction

From the conclusion shown in 4.2.2, the number of outliers is large by setting the significant level $\alpha = 0.05$ as well as $\alpha = 0.01$. Also, some points are repeatedly detected as potential outliers. Two points, the 37th and the 137th point are detected as potential outliers for all three matrices. Two potential outliers, the 207th, and the 520th point are shown in two of three matrices. In addition, as mentioned in every matrix, some points are potential outliers for two variables' outlier detections.

Since the individual t-test repeats hundreds of times, the threshold $\alpha = 0.01$ may be too large to be utilized. Hence, the Bonferroni Correction is used for summarizing conclusions. As demonstrated in 4.2.2, $\alpha^{*} = 0.00009$.

With new significant level α^{*}, two potential outliers, the 36th point, and the 242nd point, are detected for the parameter proportion and no potential outlier for the parameter guess and slip in Case I. In Case II, five points are potential outliers for the parameter guess and six points are potential outliers for the parameter proportion but no potential outlier for the parameter slip. One point, the 236th, is a potential outlier for both the parameter guess and the parameter proportion. Another point, the 242nd, is a potential outlier for the parameter proportion in Case III. In all cases, no potential outlier for the parameter slip. There are some repeated potential outliers: two points are potential outliers for both the parameter guess and the parameter proportion in Case II; one point is a potential outlier for the parameter proportion in both Case I and Case III.

5. Conclusions

From the whole process shown in Section 4, the potential outliers in dataset can be detected. The individual t-tests for every point provide more accurate and more convincing conclusions. As shown in Section 4.1, the skill "QT3" is the easiest skill for students to master and the skill "Dim3" is the most difficult skill. Also, the skills "Dim1" and "Dim2" are highly correlated while the skills "QT1" and "QT4" are the least correlated. In the outlier detection, the results of Bonferroni Correction are used to summarize the potential outliers. With the Bonferroni Correction, the number of detected potential outliers decreases and the conclusion is more convincing. In the next step of this research, other CDMs will be considered and applied.

Reference

Akaike H (1973). “Information Theory and an Extension of the Maximum Likelihood Principle.” In B Petrov, F Csake (eds.), *Second International Symposium on Information Theory*, pp. 267-281. Akademiai Kiado, Budapest.

Chen WH, Thissen D (1997). “Local Dependence Indexes for Item Pairs Using Item Response Theory.” *Journal of Educational and Behavioral Statistics*, **22**, 265-289.

de la Torre J (2011). “The Generalized DINA Model Framework.” *Psychometrika*, **76**, 179-199.

Gentle JE (2009). *Computational Statistics*.

George AC, Robitzsch A, Kiefer T, Groß J, Ünlü A (2016). “The R Package CDM for Cognitive Diagnosis Models.” *Journal of Statistical Software*, **74**(2).

Henson RA, Templin JL, Willse JT (2009). “Defining a Family of Cognitive Diagnosis Models Using Log-Linear Models with Latent Variables.” *Psychometrika*, **74**, 191-210.

James G, Witten D, Hastie T, Tibshirani R (2017). *An Introduction to Statistical Learning with Applications in R*.

Package ‘CDM’ (2018). https://cran.rproject.org/web/packages/CDM/CDM.pdf.

Schwarz G (1978). “Estimating the Dimension of a Model.” *The Annals of Statistics*, **6**, 461-464.

Tatsuoka KK (1983). “Rule Space: An Approach for Dealing with Misconceptions Based on Item Response Theory.” *Journal of Educational Measurement*, **20**, 345-354.

von Davier M (2008). “A General Diagnostic Model Applied to Language Testing Data.” *British Journal of Mathematical and Statistical Psychology*, **61**, 287-307.

Acknowledgement

I would like to express my sincerest appreciation for all people who offer me the help during the research process. First of all, I would like to appreciate my thesis instructor Dr. Gongjun Xu for the precious opportunity. I have learnt a lot from his patience as well as his erudition in Statistics. And for that, I am deeply grateful. While there is much more knowledge for me to study, working on this research helps me get better understandings and preparations for my further study. Then, I would like to appreciate my advisor, Gina Cornacchia. It is her who gives me the courage to pursue the research in Statistics. Also, I would like to thank my family, my parents, Kebin Zhao and Yuxue Sun, and the most important my dear grandmother. You are my faith and my power to accomplish my dreams. Without you, I could not be where I am. Thank you all for enduring supports all these years. Besides, I am wholeheartedly grateful for my best friends who encourage me and accompany me to challenge myself and try something new: Wanggang Shen, Xiaolei Wang, Xinying He, Xiye Wang, Yifan Sun, and Yipeng He. Finally, I would like to thank the University of Michigan, a magic place filled with opportunities and chances. I will always remember “Always Leading, Forever Valiant.” GO BLUE!