STAT 601

Final Project

NAIQING CAI

ncai5@wisc.edu

CONTENTS

1. Calories
   1. Data Preprocessing

1.1.1 Data Structure

We have the dataset about “common house food“, which includes the ingredients for each food items. So we need to first analyze which variable should be included into the model.

It includes 22 columns and 962 observations. Since there is one row is missing and I remove it from the dataset at first.

After that, we should get a general idea of the dataset, I classify the data into the following shape:

|  |  |
| --- | --- |
| Food Items |  |
| Weight (in grams) |  |
| Water (in grams) |  |
| KCal |  |
| Protein (in grams) |  |
| Cholesterol (in mg) |  |
| Carbohydrates (in grams) |  |
| Fats (in grams) | Fat  SatFat  MonoUnSatFat  PolyUnSatFat |
| Minerals (in mg) | Ca  P  Fe  K  Na |
| Vitamins | VitaA (IU)  VitaA (RE)  Thiamin (in mg)  Riboflavin (in mg)  Niacin (in mg)  VitaC (in mg) |

Table 1.1 Data Structure on Common House Food Dataset

Based on the above data classification table, I decide to include the following 8 variables in the initial model and each variable may have one or more items: Water, Weight, Protein, Carbohydrates, Cholesterol, Vitamins, Minerals, Fats.

1.1.2 Data Analysis

1.1.2.1 Multicollinearity

(1) Detecting Multicollinearity Using Variance Inflation Factors

The idea scenario is that all the variables are uncorrelated. So for explanatory variables, we should do diagnostics for multicollinearily.

The full model is in the form:



Since there are 20 variables in the model, I choose to use variance inflation factor to do diagnostics for multicollinearily.

Variance for beta(k) is defined as follows:



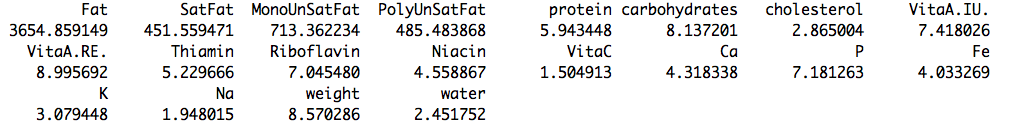


Table VIF for explanatory variables for full model

4 of them (all regarding fats items ) are even greater than 10. So, we can conclude that the multicollinearily among these 4 items will have a large impact on the inference.

Also, the mean of all the VIF values is 269.4273, which is greater than 1. So, there may be serious multicollinearily problems and we should do some transformation for the data.

(2) Transformation for Reducing Data-based Multicollinearity

In order to reduce the multicollinearily, we could opt to remove some insignificant variables out of the predictors from the model.

Alternatively, if we have a good scientific reason for needing both of the predictors to remain in the model, we could go out and collect more data and then will reduce the multicollinearily among the predictors.

For the context in this problem, I choose to use the first method and by using the stepwise function in R and with BIC+Both criterion since it better for result of model fitting. The result is below in the table:

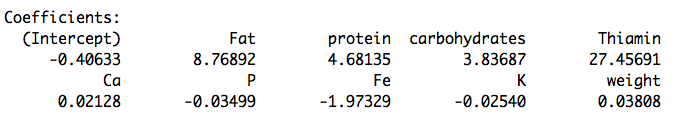


Table Coefficients for the Reduced Model

Then we remove 11 insignificant variables (VitaA.IU., VitaA.RE., Niacin, water, PolyUnSatFat, SatFat, MonoUnSatFat, cholesterol, Na, Riboflavin, VitaC) from the original full model, and the model is the following form:



Again, I use variance inflation factor to do diagnostics for multicollinearily. Then the result is in the following table.

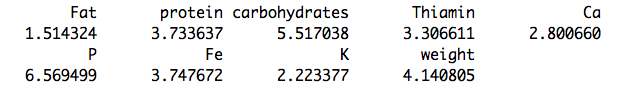


Table VIF for explanatory variables for reduced model

Then, we find that all of the item fat is smaller than 10. Also the mean of all the VIF values is 3.72818, which is smaller than the full model. The result is much better than the original data. So we can conclude that the effect of multicollinearity on the model inference is much smaller than before.

1.1.2.2 Regression Assumptions

(1) Checking Assumptions

Based on the model selected in the I take studentized residuals to check the assumptions. If the studentized residuals are large, it means that the observations have outliers and we need to do transformation on the data.

The results can be seen from the following figures:

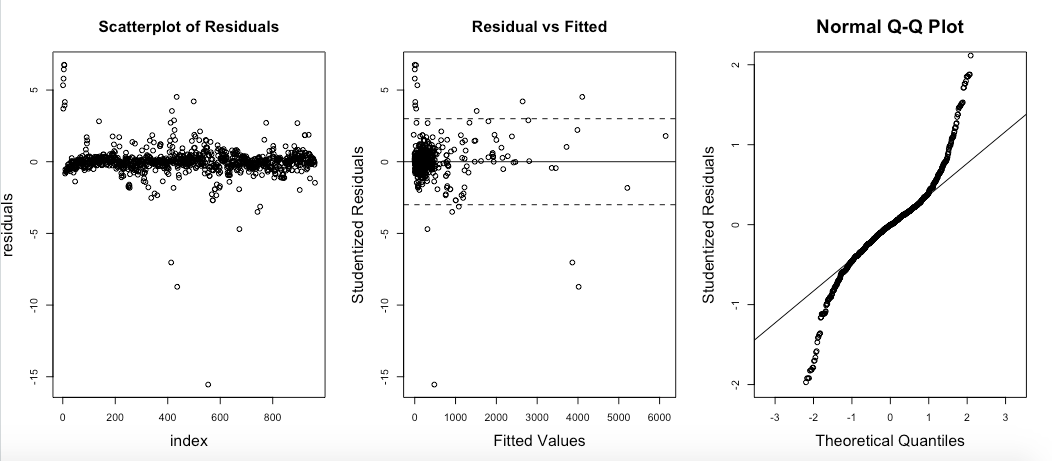


Figure Checking Regression Assumptions on Original dataset

Assumption 1: Independence



Based on the scatterplot of residuals, we can conclude that the assumption of independence is satisfied. So we do not need to do any transformation about the data.

Assumption 2: Linearity



From the data pattern, we could easily find that the model is linear regression.

Assumption 3: Equal Variance



Based on the plot of residuals v.s. fitted, we find that the variance of residuals do not have distinct structures and then we conclude that the assumption of equal variance is nearly satisfied well, so we do not need to do some transformation.

Assumption 4: Normal Distribution



Based on the normal QQ plot, we conclude that the data is not fully from normal distribution, so we need to do some transformation about the data.

So, I choose to do some transformation on the dataset and since y (calories) have some zero values, so I do square root transformation on the response variables to adjust the data

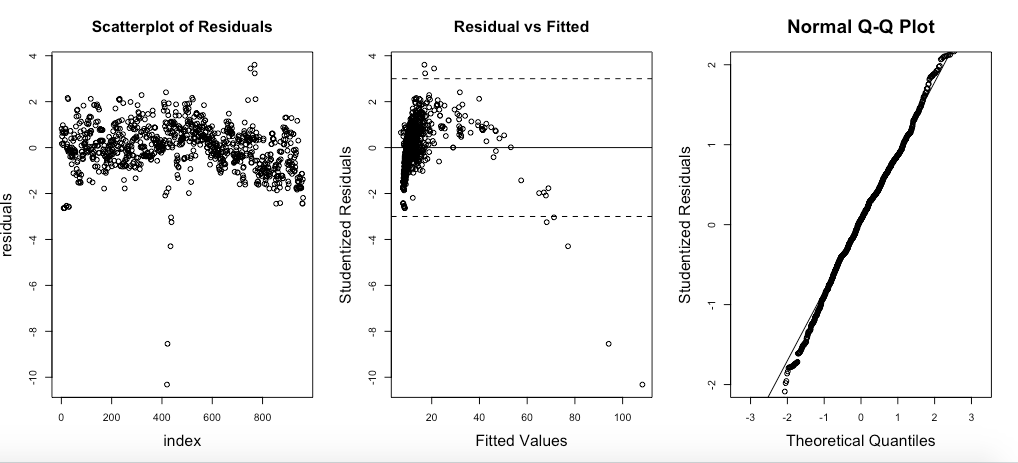


Figure Checking Regression Assumptions on Transformed dataset

Based on the plot of the transformed data, we could find that the data is much better than before, and the assumptions are nearly satisfied, especially the result of QQ plot which is totally from normal distribution.

In this way, we will use the transformed dataset to fit the model in the following steps.

1.1.2.3 Outliers

I take studentized residuals to check the assumptions. If the studentized residuals are large, it means that the observation is an outlier.

Since we find evidence in the last step that there may be some outliers in the observations, we need to identify and drop them. I use the following four methods to access it: Leverage, DFFITS, Cook’s distance, DFBETAS.

601%20final%20figure/Rplot08.pdf

(1) Leverage

(2) DFFITS

(3) Cook’s distance

(4) DFBETAS.

* 1. Model Fitting and Evaluation

We consider a model in the following term:



And then, we need to come up with a procedure to obtain the best model.

(1) Stepwise: Backward+AIC

First, we use stepwise method to choose the best model since the number of predictors are large and stepwise will be more efficient for the model. The working theorem is by using backward and AIC criterion to eliminate some of the predictors until all the predictors in the model can not be delated.

We start from the full model and delete explanatory variables. Once deleted, an explanatory variable can not be added to the model again.

We use the step() function in R to achieve the goal.

Step by step, and finally we get the result as follows:

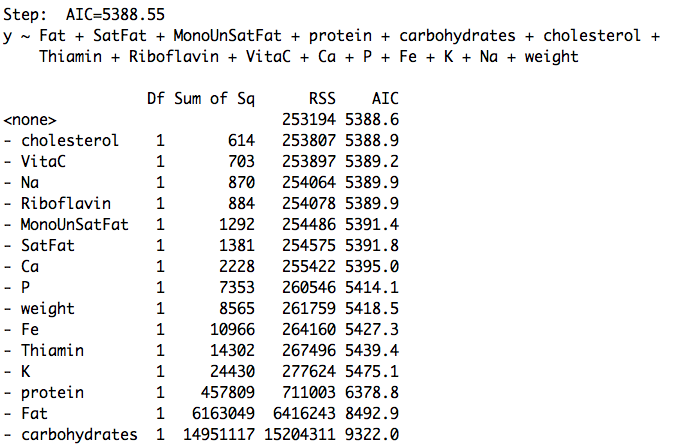


Table Backward+AIC Elimination Step

Based on the table, we can find that some of the variables have been removed under the constrains. But there are still too more predictors in the model, so we need to move on.

(2) Elimination based on the Deviances

After the stepwise, then we need to use another method to remove more predictors. I choose the way based on the deviances. We can judge whether the model is adequate or not by goodness-of-test.

Goodness of test is based on the relationship between the deviance and degree of freedom.

* 1. Prediction

In order to test my model’s prediction power, I need to use the new data to predict and get it’s performance.

I use the following new dataset, putting them into the model.

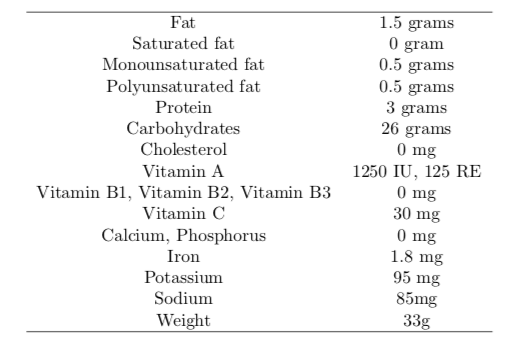


Table X value for Prediction

1. Audibility
   1. Data Preprocessing

2.1.1 Data Structure

The whole data includes 6 columns, and the total observations are 672. The following is a code book for the dataset.

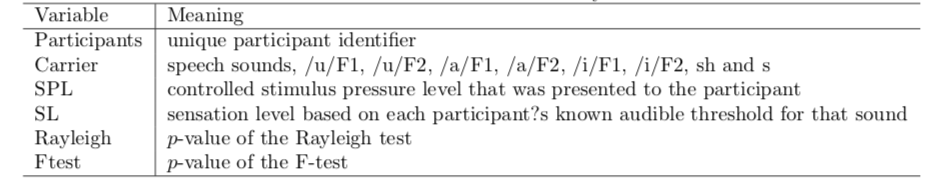


Table Code Book for the Audibility Dataset

2.1.2 Getting Categorical Data into Shape

Since “Carrier”, “SPL”, “SL” are all binary features, so I do a transformation to get them into 0-1 form so that the step of regression can be much more easily. (details in appendix)

* 1. Experiments on Accuracy of EFRs

2.2.1 General Analysis on Accuracy of EFRs on Predicting Audibility of Speech Stimulus

To make a general analysis of the accuracy of EFRs of predicting the audibility of speech stimulus, we can use the table about types of error and statistical power.

1. F-test

**Analysis:**

We can judge the accuracy of EFRs, based on the F-test.

First, we calculate the percentage of true label v.s true predicted label. We need to find the number of SL that is larger than zero, and also the number of Ftest which is smaller than 0.05, representing the “detected” EFRs. We find that the number of SL that is larger than zero is 560. And the number of Ftest which is smaller than 0.05 is 355. So the probability that the predicted audibility by EFRs divided by the true audibility for each stimulus for participate is 0.6339286.

Then, we calculate the percentage of true label v.s false predicted label. We find that the number of SL that is larger than zero is 560. And the number of Ftest which is larger than 0.05 is 205, representing the “undetected” EFRs. So in this way, the probability that the predicted inaudibility by EFRs divided by the true audibility for each stimulus for participate is 0.3660714. That means type one error is 0.3660714.

Next, we calculate the percentage of false label v.s true predicted label. We find that the number of SL that is smaller than zero is 112, which means the stimuli were inaudible. And the number of Ftest which is smaller than 0.05 is 5, representing the “detected” EFRs. In this way, the probability that the predicted audibility by EFRs divided by the inaudibility for each stimulus for participate is 0.04464286, which means that type two error is 0.04464286.

Finally, we calculate the percentage of false label v.s false predicted label. We find that the number of SL that is smaller than zero is 112, which means the stimuli were inaudible. And the number of Ftest which is larger than 0.05 is 107, representing the “undetected” EFRs. In this way, the probability that the predicted inaudibility by EFRs divided by the inaudibility for each stimulus for participate is 0.9553571, which means the power of F-test is 0.9553571.

**Conclusions：**

Based on the following table, we can see that F-test has a high accuracy of 0.9553571 to predict on the inaudible stimulus, but a lower accuracy of 0.6339286 to predict on audible stimulus.

|  |  |  |
| --- | --- | --- |
| predicted  reality | True(Audible) | FALSE(inaudible) |
| TRUE(audible) | 0.6339286 | 0.3660714 |
| Flase(inaudible) | 0.04464286 | 0.9553571 |

Table F-test Accuracy

1. Rayleigh-test

**Analysis:**

Also, we can judge the accuracy of EFRs, based on the Rayleigh-test.

First, we calculate the percentage of true label v.s true predicted label. We need to find the number of SL that is larger than zero, and also the number of Rayleigh-test which is smaller than 0.05, representing the “detected” EFRs. We find that the number of SL that is larger than zero is 560. And the number of Rayleigh-test which is smaller than 0.05 is 380. So the probability that the predicted audibility by EFRs divided by the true audibility for each stimulus for participate is 0.6785714.

Then, we calculate the percentage of true label v.s false predicted label. We find that the number of SL that is larger than zero is 560. And the number of Rayleigh-test which is larger than 0.05 is 180, representing the “undetected” EFRs. So in this way, the probability that the predicted inaudibility by EFRs divided by the true audibility for each stimulus for participate is 0.3214286. That means type one error is 0.3214286.

Next, we calculate the percentage of false label v.s true predicted label. We find that the number of SL that is smaller than zero is 112, which means the stimuli were inaudible. And the number of Rayleigh-test which is smaller than 0.05 is 10, representing the “detected” EFRs. In this way, the probability that the predicted audibility by EFRs divided by the inaudibility for each stimulus for participate is 0.08928571, which means that type two error is 0.08928571.

Finally, we calculate the percentage of false label v.s false predicted label. We find that the number of SL that is smaller than zero is 112, which means the stimuli were inaudible. And the number of Rayleigh-test which is larger than 0.05 is 102, representing the “undetected” EFRs. In this way, the probability that the predicted inaudibility by EFRs divided by the inaudibility for each stimulus for participate is 0.9107143, which means the power of F-test is 0.9107143.

**Conclusions：**

Based on the following table, we can see that Rayleigh-test has a high accuracy of 0.9107143 to predict on the inaudible stimulus, but a lower accuracy of 0.6785714 to predict on audible stimulus.

|  |  |  |
| --- | --- | --- |
| predicted  reality | True(Audible) | FALSE(inaudible) |
| TRUE(audible) | 0.6785714 | 0.3214286 |
| Flase(inaudible) | 0.08928571 | 0.9107143 |

Table Rayleigh-test Accuracy

1. F-test and Rayleigh-test

**Analysis:**

Finally, we can judge the accuracy of EFRs, based on both F-test and Rayleigh-test.

First, we calculate the percentage of true label v.s true predicted label. We need to find the number of SL that is larger than zero, and also the number of both F-test and Rayleigh-test which are smaller than 0.05, representing the “detected” EFRs. We find that the number of SL that is larger than zero is 560. And the number of both F-test and Rayleigh-test which are smaller than 0.05 is 343. So the probability that the predicted audibility by EFRs divided by the true audibility for each stimulus for participate is 0.6125.

Then, we calculate the percentage of true label v.s false predicted label. We find that the number of SL that is larger than zero is 560. And the number of both F-test and Rayleigh-test which are larger than 0.05 is 217, representing the “undetected” EFRs. So in this way, the probability that the predicted inaudibility by EFRs divided by the true audibility for each stimulus for participate is 0.3875. That means type one error is 0.3875.

Next, we calculate the percentage of false label v.s true predicted label. We find that the number of SL that is smaller than zero is 112, which means the stimuli were inaudible. And the number of both F-test and Rayleigh-test which are smaller than 0.05 is 3, representing the “detected” EFRs. In this way, the probability that the predicted audibility by EFRs divided by the inaudibility for each stimulus for participate is 0.02678571, which means that type two error is 0.02678571.

Finally, we calculate the percentage of false label v.s false predicted label. We find that the number of SL that is smaller than zero is 112, which means the stimuli were inaudible. And the number of both F-test and Rayleigh-test which are larger than 0.05 is 109, representing the “undetected” EFRs. In this way, the probability that the predicted inaudibility by EFRs divided by the inaudibility for each stimulus for participate is 0.9732143, which means the power of F-test is 0.9732143.

**Conclusions：**

Based on the following table, we can see that combination of F-test and Rayleigh-test has the highest accuracy of 0.9732143 to predict on the inaudible stimulus, but also the lowest accuracy of 0.6125 to predict on audible stimulus.

|  |  |  |
| --- | --- | --- |
| predicted  reality | True(Audible) | FALSE(inaudible) |
| TRUE(audible) | 0.6125 | 0.3875 |
| Flase(inaudible) | 0.02678571 | 0.9732143 |

Table Combination of F-test and Rayleigh-test Accuracy

2.2.2 Relationship between Accuracy of EFRs on Predicting Audibility of Speech Stimulus and Carriers or Frequency Groups

To explore the accuracy of EFRs on predicting audibility of each speech stimulus differ between carriers and frequency groups, we need to do some regression on the data.

We mainly consider the detectability (EFRs), which is a binary outcome, so I choose to use logistic regression since it fits the condition that response value is a binary one.

* + - 1. One Factor Model

1. Model

We consider the following two one factor regression models:



1. Arguments and Parameters:

EFRs: representing the detectability of the stimulus

SPL: controlled stimulus pressure level that was presented to the participant, which isa categorical data with 4 levels

Carrier: speech sounds, which is a categorical data with 8 levels

Beta0: intercept for detectability of the stimulus

Beta(i): difference in detectability of the stimulus between i level of SPL and reference level of SPL

Beta(j): difference in detectability of the stimulus between j level of carrier and reference level of carrier

1. Analysis

For model (1)

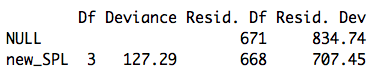


Table Deviance of One Factor Model (1)

Null model has a deviance of 834.74 on 671 degrees of freedom, the p-value is nearly asymptotic to 0, which doesn’t pass the goodness-of-test, so we reject the null hypothesis that all detectability is the same with the change of SPL.

Introducing SPL variable leads to substantial reduction of 127.29 deviance at only 3 degrees of freedom. So the variable of SPL has significant effect on the detectability of the stimulus.

For model (2)

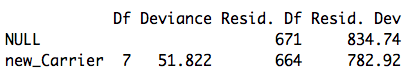


Table Deviance of One Factor Model (2)

Also, null model doesn’t pass the goodness-of-test. Introducing Carrier variable leads to substantial reduction of 127.29 deviance at only 7 degrees of freedom. So the variable of Carrier has significant effect on the detectability of the stimulus.

* + - 1. Two Factor Model

1. Additive Model



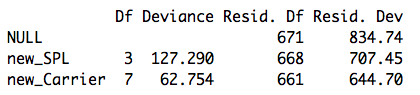


Table Deviance of Two Factor Model

Beta0: intercept for detectability of the stimulus

Beta(i): difference in detectability of the stimulus between i level of SPL and reference level of SPL

Alpha(j): difference in detectability of the stimulus between j level of carrier and reference level carrier

The additive model has a deviance of 190.04 at only 10 degrees of freedom. So the model provides a good description of the data.

1. Two Factor Model with Interaction



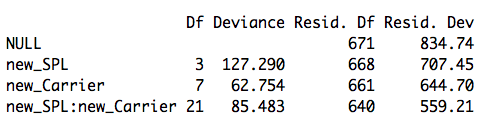


Table Deviance of Two Factor Model with Interaction

The two factor model with interaction has a deviance of 275.53 at only 31 degrees of freedom. So the interaction between SPL and carrier has a significant effect on the model and thus the model provides a good description of the data.

* + - 1. Conclusions

Based on the analysis above, we can get the following table.

|  |  |  |
| --- | --- | --- |
| Model | Deviance | d.f. |
| Null | 834.74 | 671 |
| *One Factor Model* |  |  |
| SPL | 707.45 | 668 |
| Carrier | 782.92 | 664 |
| *Two Factor Model* |  |  |
| SPL+Carrier | 644.7 | 661 |
| SPL+Carrier+ SPL\*Carrier | 559.21 | 640 |

Table Deviance for the Whole Model

Our conclusion is that the model with two factors and interaction is the most accurate model so that accuracy differs both between carriers and frequency groups. What’s more, accuracy differs between the interaction of carriers and frequency groups, too.

* 1. Performance between the F-test and the Rayleigh in Predicting Audibility

2.3.1 F-test in Predicting Audibility

Based on the results of (2.2), we can get the table as follows:

|  |  |  |
| --- | --- | --- |
| predicted  reality | True(Audible) | FALSE(inaudible) |
| TRUE(audible) | 0.6339286 | 0.3660714 |
| Flase(inaudible) | 0.04464286 | 0.9553571 |

Table F-test Accuracy

2.3.2 Rayleigh in Predicting Audibility

Based on the results of (2.2), we can get the table as follows:

|  |  |  |
| --- | --- | --- |
| predicted  reality | True(Audible) | FALSE(inaudible) |
| TRUE(audible) | 0.6785714 | 0.3214286 |
| Flase(inaudible) | 0.08928571 | 0.9107143 |

Table Rayleigh-test Accuracy

2.3.3 Combination of F-test and Rayleigh in Predicting Audibility

Based on the results of (2.2), we can get the table as follows:

|  |  |  |
| --- | --- | --- |
| predicted  reality | True(Audible) | FALSE(inaudible) |
| TRUE(audible) | 0.6125 | 0.3875 |
| Flase(inaudible) | 0.02678571 | 0.9732143 |

Table Combination of F-test and Rayleigh-test Accuracy

2.3.4 Conclusions

Making a comparison of the two tables, we draw the following conclusions:

1. Rayleigh-test has higher accuracy of predicting audibility on the audible stimulus.
2. F-test has higher accuracy of predicting audibility on the inaudible stimulus.
3. If we combine the results of F-test and Rayleigh-test, then the accuracy of prediction on both audible and inaudible stimulus will be higher than based on one test.
4. The combination of two test will make the accuracy of predicting audibility on the inaudible stimulus higher, but accuracy of predicting audibility on the audible stimulus lower.
5. Once we have a higher accuracy of predicting audibility on the inaudible stimulus, accuracy of predicting audibility on the audible stimulus will become lower. They are negative related.

2.4 Minimum SL for Detectability

2.4.1 Global Minimum of SL

To find the minimum of SL needed in order for the EFR to detect a response, we should first search the observations that the value of SL is larger than 0. Then we find there are 560 observations. Next, we based on different tests to find the minimum of SL. The results are in the following table.

|  |  |
| --- | --- |
| Method | Minimum of SL |
| F-test | 1.36 |
| Rayleigh-test | 3.46 |
| F-test & Rayleigh-test | 4.64 |

Table Global Minimum of SL

We find that based on the F-test the minimum 1.36 is the smallest. On the Rayleigh-test, the minimum is 3.46. And, on the combination of two tests, the minimum 4.64 is largest

2.4.2 Relationship between Minimum of SL and Carrier or Frequency Groups

2.4.2.1 Carrier

|  |  |  |
| --- | --- | --- |
| Method | Carrier | Minimum of SL |
| F-test | a\_F1 | 4.9 |
|  | a\_F2 | 6.71 |
|  | i\_F1 | 3.46 |
|  | i\_F2 | 6.36 |
|  | u\_F1 | 13.66 |
|  | u\_F2 | 6.2 |
|  | s | 4.64 |
|  | sh | 5.63 |
| Rayleigh-test | a\_F1 | 4.9 |
|  | a\_F2 | 6.71 |
|  | i\_F1 | 8.46 |
|  | i\_F2 | 1.36 |
|  | u\_F1 | 8.66 |
|  | u\_F2 | 6.2 |
|  | s | 4.64 |
|  | sh | 5.63 |
| F-test & Rayleigh-test | a\_F1 | 4.9 |
|  | a\_F2 | 6.71 |
|  | i\_F1 | 8.46 |
|  | i\_F2 | 6.36 |
|  | u\_F1 | 13.66 |
|  | u\_F2 | 11.2 |
|  | s | 4.64 |
|  | sh | 5.63 |

Table Relationship between Minimum of SL and Carrier

Based on the above table, I use three different colors to differ three levels of carrier: low, mid and high frequencies. The F1 carriers are low frequency dominant, the F2 carriers are mid frequency dominant and the fricatives (sh and s) are high frequency dominant.

In general, we can find that with the higher carrier, and will have the smallest minimum of SL. The lower the carrier is, the larger the minimum of SL.

Also, we can conclude that among F1 and F2 level, there are three different forms: “a”, “i”, and “u”. We find that for F1 level, these three kinds have a distinct difference in minimum of SL than that in F2 level. And “u” always has the largest minimum of SL, “a” always has the smallest minimum of SL.

2.4.2.2 Frequency Groups

In order to have a more detailed exploration of the problem how does the minimum SL vary by frequency group, we need to classify the carrier feature into three categories: low, mid and high.

Then we calculate the minimum of SL of each frequency again based on three different methods.

|  |  |  |
| --- | --- | --- |
| Method | Frequency Groups | Minimum of SL |
| F-test | low |  |
|  | mid |  |
|  | high |  |
| Rayleigh-test | low |  |
|  | mid |  |
|  | high |  |
| F-test & Rayleigh-test | low |  |
|  | mid |  |
|  | high |  |

Table Relationship between Minimum of SL and Frequency Groups

2.5 Limitation and Improvement