STAT 601

Final Project

NAIQING CAI

ncai5@wisc.edu

Contents

[1. Calories 3](#_Toc532741892)

[1.1 Data Preprocessing 3](#_Toc532741893)

[1.2 Model Fitting 9](#_Toc532741894)

[1.3 Prediction 10](#_Toc532741895)

[2. Audibility 11](#_Toc532741896)

[2.1 Data Preprocessing 11](#_Toc532741897)

[2.2 Experiments on Accuracy of EFRs on Predicting Audibility of Speech Stimulus 11](#_Toc532741898)

[2.3 Minimum SL for Detectability 17](#_Toc532741899)

[2.4 Limitation and Improvement 20](#_Toc532741900)

[3. Electro-chemical 23](#_Toc532741901)

[3.1 Conservative and Additive Matrix 23](#_Toc532741902)

[3.2 Least Square Estimate 24](#_Toc532741903)

[3.3 Model Fittting 25](#_Toc532741904)

[3.4 Transformation 25](#_Toc532741905)

# 1. Calories

*Summary*

In this calories context, our goal is to obtain the “best” model by which we can get the most accurate prediction for a given value of X. Thus, I divide my report into 3 parts: data preprocessing, model fitting and model prediction.

For data preprocessing, I deal with some problems including: multicollinearity, checking model assumptions, influential observations and outliers.

For model fitting, I choose the model under some constraints of the form (1) and conclude that the best model is like the form (2).





For prediction, I gain a new set of data and predict the calories of it.

## 1.1 Data Preprocessing

1.1.1 Data Structure and Analysis

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 Diagnostics

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:



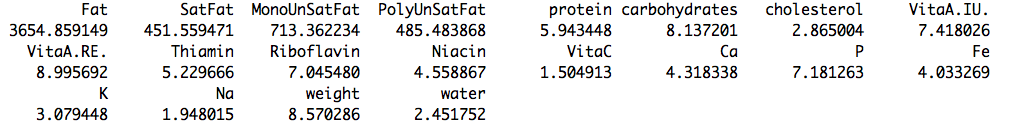


Table 1.2 VIF for explanatory variables for full model

We find that 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:

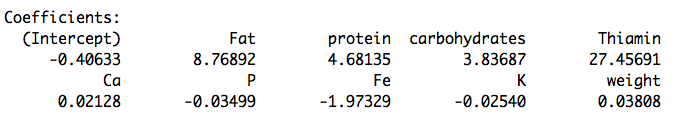


Table 1.3 Coefficients for the Reduced Model

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 reduced model is the following form:



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

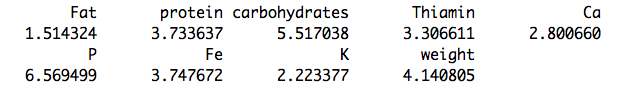


Table 1.4 VIF for Explanatory Variables for Reduced Model

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) Diagnostic Plots

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:

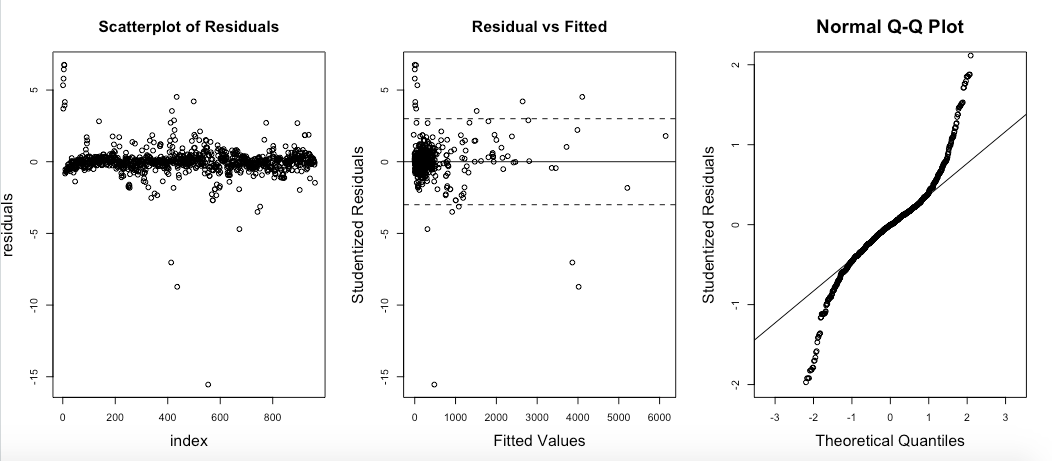


Figure 1.5 Checking Regression Assumptions on Original dataset

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 may need to do some kinds transformation about the data.

(2) Boxcox Checking

I choose to use boxcox to judge if I indeed need to do some transformation on the dataset.

By doing boxcox on the reduced model, we can easily find based on the following figure that lambda is 1 and it means that we do not need to do any transformation of powerful number. I think maybe it’s because the model is not so sensitive to the assumption of normal distribution that we do not need to completely access this assumption.

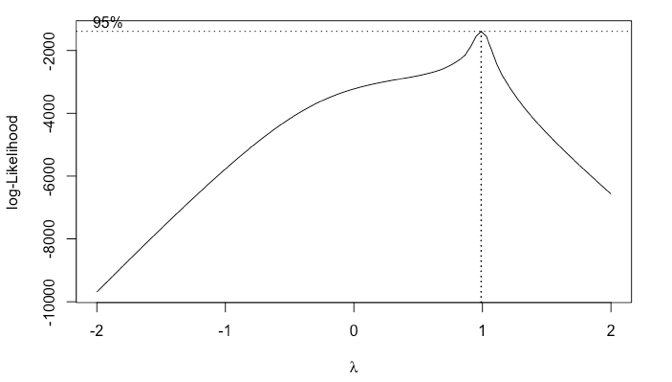


Figure 1.6 Boxcox of for Reduced Model

1.1.2.3 Measures of influence

I take studentized residuals to check the influence. If the studentized residuals are large, it means that the observations are influential points. I use the following four methods to access it: Leverage, DFFITS, Cook’s distance, DFBETAS. We can all four plots shows that there may be some influential observations in the data. They may not all be outliers, so we need to check further.

601%20final%20figure/Rplot08.pdf

Figure 1.7 Identification of Influential Observations

1.1.2.4 Outliers

In order to find the outliers, we plot the reduced model and we can identify from the figure below that the observation of 413, 436, 554 are three outliers. So we need to drop these three data point as the first step.

Rplot.pdf

Figure 1.8 Plot of Reduced Model

After dropping these three outliers, we fit the new reduced model with new dataset and plot the model again. I find that this time the outlier is not so extinct. So we stop and use these data to continue the analysis.

Rplot03.pdf

Figure 1.9 Plot of Reduced Model without Outliers

## 1.2 Model Fitting

We consider a model in the following term:



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

I choose to use the best subsets method with various model selection criteria. A best subsets procedure identifies a group of subset models that gives the best values of a given criterion. Since the number of explanatory variables in this context is large, all possible best best subsets may not be feasible. In this case, a stepwise selection procedure offers a feasible approach.

I use “myregsub.R” to as a procedure to choose the best model under the constraint of the model with seven explanatory variables.

Then, based on the result, I find that the best model which fits the requirements includes the following seven variables:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Number of variables | Weight  Protein  Fat  Chol  Carb  K  Thiamin | rsq | rss | adjr2 | cp | bic |
| 7 | 0.9990175 | 278026.0 | 0.9990103 | 92.18697 | -6600.344 |

Table 1.10 Best Subsets Procedure

The final model now is like this way:



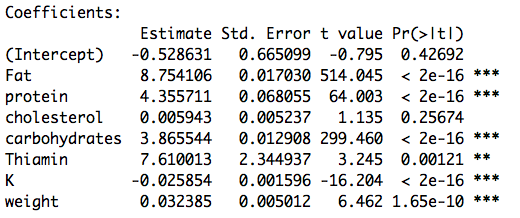


Table 1.11 R Summary of Final Model

And also we can find that multiple R-squared is 0.9993, which means that this model can perfectly explain 99.93% of the variables, which will lead to a highly accurate prediction for a given value of X.

## 1.3 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 gain the following new dataset:

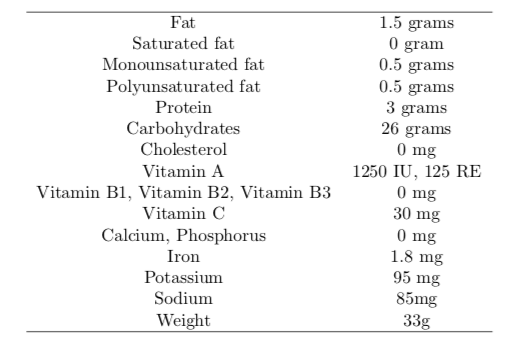


Table X value for Prediction

Since my final model only includes 7 variables: Weight, Protein, Fat, Chol, Carb, K, Thiamin. I just put these 7 into the final model.

Then, I get the result as follows:

(1) Predictive Value



(2) Predictive Interval





# 2. Audibility

## 2.1 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.

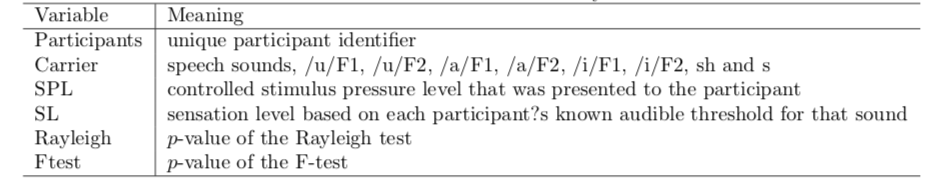


Table Code Book for the Audibility Dataset

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)

## 2.2 Experiments on Accuracy of EFRs on Predicting Audibility of Speech Stimulus

2.2.1 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 or 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.

2.2.2.1 Carriers

2.2.2.1.1 Main Effect Analysis

(1) Model

We consider the following one factor regression models:



(2) Arguments and Parameters

|  |  |
| --- | --- |
| EFRs | representing the detectability of the stimulus |
| Carrier | speech sounds, which is a categorical data with 8 levels |
| Beta0 | accuracy of EFRs (detectability of the stimulus) for reference level |
| Beta(j) | difference in accuracy of EFRs (detectability of the stimulus) between j level of carrier and reference level of carrier |
| Beta0+ Beta(j) | accuracy of EFRs (detectability of the stimulus) at j level of carrier |

Table Arguments and Parameters for the Model

(3) Analysis

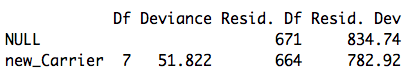


Table Anova Table for One Factor Model

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.

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 and accuracy of EFRs.

2.2.2.1.2 Effect of Eight Levels of Carriers on Accuracy of EFRs

In order to explore the effect of 8 levels in carriers on accuracy of EFRs, we need to fit the following model both the result of F-test and Rayleigh-test.



(1) F-test

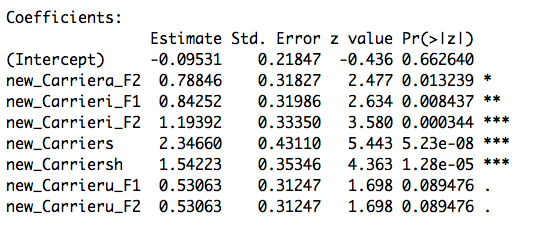


Table Model Fitting for EFRs~Carriers on F-test



Based on the result of model fitting for EFRs~Carriers on F-test above, we conclude the p-values for each 8 level in carriers so that we can judge the significance of each level in the model.

I find that all of the levels are significant at 90% confidence level. And “i\_F2”, “s”, “sh” is the three most significant effects in the model.

|  |  |  |
| --- | --- | --- |
| level of carriers | Logit(pi) | pi |
| a\_F1 | -0.09531 | 0.4761905 |
| u\_F1 | 0.43532 | 0.6071433 |
| i\_F1 | 0.74721 | 0.6785705 |
| a\_F2 | 0.69315 | 0.6666673 |
| u\_F2 | 0.43532 | 0.6071433 |
| i\_F2 | 1.09861 | 0.7499996 |
| s | 2.25129 | 0.9047617 |
| sh | 1.44692 | 0.809524 |
| Average | / | 0.6875 |

Table Values of pi of Carriers on F-test

Based on the values of pi on F-test above, we can conclude that different levels of carriers will lead to a difference in accuracy in predicting the audibility of stimulus. The level “s” will have the highest accuracy and “a\_F1” will have lowest accuracy. And the average accuracy of all levels on F-test is 0.6875.

(2) Rayleigh-test

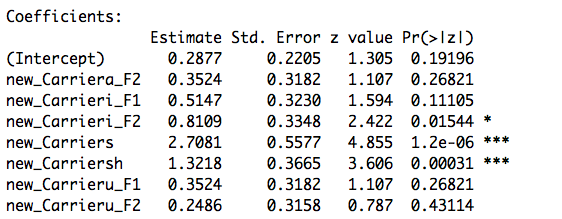


Table Model Fitting for EFRs~Carriers on Rayleigh-test

Based on the result of model fitting for EFRs~Carriers on Rayleigh-test above, we conclude p-values for 3 out of 8 levels are significant in the model, and they are: “i\_F2”, “s”, “sh”.

|  |  |  |
| --- | --- | --- |
| level of carriers | Logit(pi) | pi |
| a\_F1 | 0.2877 | 0.571433 |
| u\_F1 | 0.6400 | 0.6547535 |
| i\_F1 | 0.8023 | 0.6904663 |
| a\_F2 | 0.6400 | 0.6547535 |
| u\_F2 | 0.5363 | 0.6309513 |
| i\_F2 | 1.0986 | 0.7499977 |
| s | 2.9957 | 0.9523795 |
| sh | 1.6094 | 0.8333281 |
| Average | / | 0.71726 |

Table Values of pi of Carriers on Rayleigh-test

Based on the values of pi on Rayleigh-test above, we can conclude that different levels of frequency will also lead to a difference in accuracy in predicting the audibility of stimulus. The level “s” has the highest accuracy and “a\_F1” will have lowest accuracy. And the average accuracy of all levels on Rayleigh-test is 0.71726.

(3) Comparison

Compared with the result on the F-test, we can find that “i\_F2”, “s”, “sh” these 3 levels always have very small p-values based on the both two testing methods: F-test and Rayleigh-test, which means they are always the most significant effects in the model. Also, based on the Rayleigh-test, the average accuracy is higher than that on F-test.

From my perspective, the small difference between of two tests may caused by the refined classification on the carrier features.

2.2.2.2 Frequency Groups

(1) Model

In order to explore the difference of accuracy between frequency groups, we need to do a transformation for the dataset. Since the data has three levels of frequency: low, mid and high, we need to classify the Carrier feature into these three categories. Then, we can fit the model as follows again.



(2) Arguments and Parameters

|  |  |
| --- | --- |
| EFRs | representing the detectability of the stimulus |
| Frequency | frequency of sounds, which is a categorical data with 3 levels |
| Beta0 | accuracy of EFRs (detectability of the stimulus) for reference level |
| Beta(j) | difference in accuracy of EFRs (detectability of the stimulus) between j level of frequency and reference level of frequency |
| Beta0+ Beta(j) | accuracy of EFRs (detectability of the stimulus) at j level of frquency |

Table Arguments and Parameters for the Model

(3) Analysis

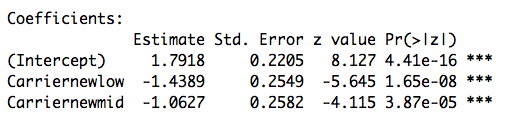


Table Model Fitting for EFRs~frequency on F-test

Based on the results of model fitting for EFRs~ frequency on F-test above, we conclude that these three levels all have significant effect on the accuracy of EFRs. The high frequency will lead to the highest accuracy of EFRs. The low frequency will lead to the lowest accuracy of EFRs.

|  |  |  |
| --- | --- | --- |
| level of frequency | Logit(pi) | pi |
| low | 0.3528 | 0.5872964 |
| mid | 0.7291 | 0.6746077 |
| high | 1.7918 | 0.8571478 |
| average | / | 0.70635 |

Table Values of pi of Frequency on F-test

Based on the table above, we can make sure that higher the frequency, the higher the accuracy. And the average accuracy is 0.70635.

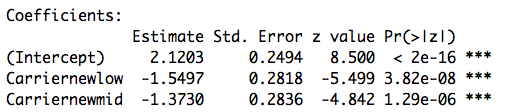


Table Model Fitting for EFRs~frequency on Rayleigh-test

Based on the results of model fitting for EFRs~ frequency on Rayleigh-test above, we also conclude that these three levels all have significant effect on the accuracy of EFRs based on the small p-values. The higher the frequency is, the higher the accuracy of EFRs will be.

|  |  |  |
| --- | --- | --- |
| level of frequency | Logit(pi) | pi |
| low | 0.5705 | 0.6388785 |
| mid | 0.7472 | 0.6785683 |
| high | 2.1203 | 0.8928606 |
| average | / | 0.7368 |

Table Values of pi of Frequency on Rayleigh-test

Based on the table above, we can make sure that higher the frequency, the higher the accuracy. And the average accuracy is 0.7368.

(3) Comparison

Comparing two testing methods, the results are almost the same, and it may be caused by the robustness of classification of frequency. But the average of accuracy on Rayleigh-test is still a little bit higher than that on F-test.

2.2.2 Performance between the F-test and the Rayleigh in Predicting Audibility

To make a general analysis of the accuracy performance of EFRs of predicting the audibility of stimulus between the F-test and the Rayleigh, we can use the table about types of error and statistical power respectively on F-test and Rayleigh.

2.2.2.1 F-test in Predicting Audibility

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

Table F-test Accuracy

2.2.2.2 Rayleigh in Predicting Audibility

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

Table Rayleigh-test Accuracy

2.2.2.3 Combination of F-test and Rayleigh in Predicting Audibility

|  |  |  |
| --- | --- | --- |
| 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

2.2.2.4 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.

## 2.3 Minimum SL for Detectability

2.3.1 Global Minimum of SL

To find the minimum of SL needed in order for the EFR to detect a response, we should first fit the following two models respectively on F-test and Rayleigh-test:





1. F-test

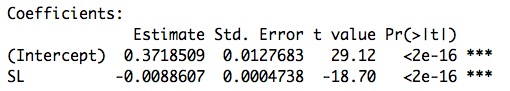


Table Global Minimum of SL on F-test

1. Rayleigh-test

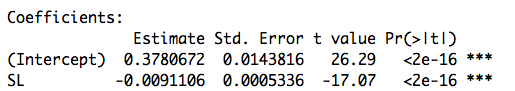


Table Global Minimum of SL on Rayleigh-test

1. Conclusions

|  |  |
| --- | --- |
| method | Minimum of SL |
| f-test | 37.02498 |
| Rayleigh-test | 36.0094 |

Table Comparison between F-test and Rayleigh-test

We can see that minimum of SL on Rayleigh-test is a bit smaller than that on F-test.

2.3.2 Relationship between Minimum of SL and Carrier or Frequency Groups

To find the relationship between minimum of SL needed in order for the EFR to detect a response and Carrier, we should fit the following two models respectively on F-test and Rayleigh-test:





* + - 1. Carrier

1. F-test

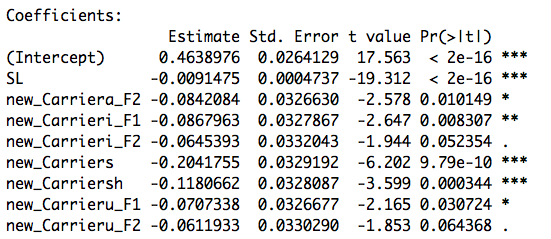


Table Relationship between Minimum of SL and Carrier on F-test

1. Rayleigh-test

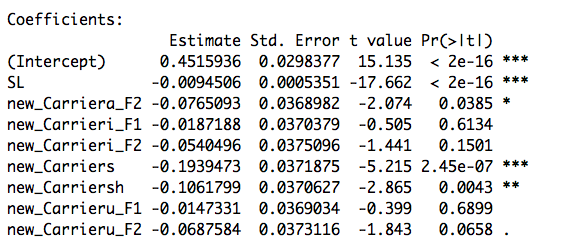


Table Relationship between Minimum of SL and Carrier on Rayleigh-test

1. Conclusions

Based on the above results of two tests, we can draw a table to find the minimum of SL based on different levels of Carrier:

|  |  |  |
| --- | --- | --- |
| Method | Carrier | Minimum of SL |
| F-test | a\_F1 | 45.24707 |
|  | a\_F2 | 36.04145 |
|  | i\_F1 | 35.75854 |
|  | i\_F2 | 38.19167 |
|  | u\_F1 | 37.51449 |
|  | u\_F2 | 38.55745 |
|  | s | 22.92671 |
|  | sh | 32.34014 |
| average of f-test |  | **35.8222** |
| Rayleigh-test | a\_F1 | 42.49398 |
|  | a\_F2 | 34.39828 |
|  | i\_F1 | 40.51329 |
|  | i\_F2 | 36.77481 |
|  | u\_F1 | 40.93503 |
|  | u\_F2 | 35.21843 |
|  | s | 21.97176 |
|  | sh | 31.25872 |
| average of Rayleigh-test |  | **35.4455** |

Table Comparison of Carrier between F-test and Rayleigh-test

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.

Finally, the average minimum of SL on Rayleigh-test is still a little bit smaller than that on F-test.

* + - 1. Frequency Groups

To find the relationship between minimum of SL needed in order for EFR to detect a response and frequency groups, we should fit the following two models respectively on F-test and Rayleigh-test:





1. F-test

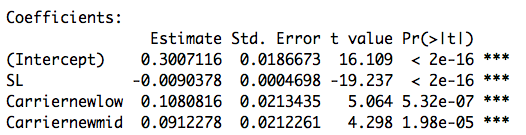


Table Relationship between Minimum of SL and Frequency Groups on F-test

1. Rayleigh-test

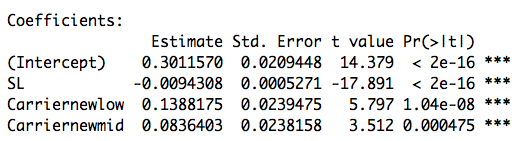


Table Relationship between Minimum of SL and Frequency Groups on Rayleigh-test

1. Conclusions

Based on the above results of two tests, we can draw a table to find the minimum of SL based on different levels of frequency groups:

|  |  |  |
| --- | --- | --- |
| Method | Frequency Groups | Minimum of SL |
| F-test | low | 39.69917 |
|  | mid | 37.83436 |
|  | high | 27.74034 |
| Rayleigh-test | low | 41.35116 |
|  | mid | 35.50042 |
|  | high | 26.63157 |

Table Comparison of Frequency Groups between F-test and Rayleigh-test

Based on the result in the table above, we find that based on the both of two testing, lowest level of frequency will always lead to the largest minimum of SL since in that way the criterion of finding the minimum of SL is stricter than other two.

## 2.4 Limitation and Improvement

2.4.1 Limitation

For the model in 2.2 as the following form, we should check the regression assumption that error terms are independent with each other and if the model is adequate.



2.4.2 Remedies

We fit additive models in the following forms

1. Additive Model



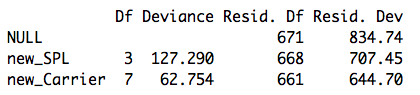


Table Deviance of Two Factor Model

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



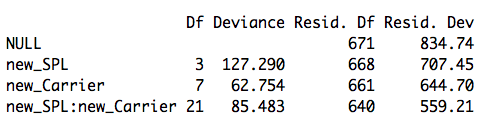


Table Deviance of Two Factor Model with Interaction

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.

|  |  |
| --- | --- |
| 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 |
| Gamma(I,j) | difference in detectability of the stimulus between the interaction of j level of carrier and i level of SPL and the reference level of interaction |

Table Parameters for the Additive Model

2.4.3 Conclusions

|  |  |  |
| --- | --- | --- |
| 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 and adequate model.

# 3. Electro-chemical

## 3.1 Conservative and Additive Matrix

For each conservative matrix V, we have:



This means that V is skew-symmetric.

On the other hand, we can conclude that not all skew-symmetric matrices are conservative, here is an example:





So not all skew-symmetric matrices are conservative.

We define a skew-symmetric matrix of the form V (i, j) = αi − αj is called additive.

We can conclude that every additive matrix is conservative:





So, V is conservative.

We can also conclude that every conservative matrix V is additive:







Finally, the space of 6×6 conservative matrices had dimension 5 and this is because the matrix is constructed by {V (1, 2), V (1, 3), V (1, 4), V (1, 5), V (1, 6)}. All the remaining parts can be calculated as a linear combination of these five values: V(u,j) = V(u,1)+V(1,j) = −V(1,u)+V(1,j). So, this will be the least components.

## 3.2 Least Square Estimate

For a single k × k table, obtain an expression for the least-squares estimate of α. Use this formula to compute αˆ for each of the three electrolytes. Explain why (α1, · · · , α5) and (α1, · · · , α5) + (c, c, c, c, c) are equivalent as parameter points in the model.

We can set this up as a typical least squares problem by transforming the k × k matrix Y into a

k2 × 1 vector y. Then, if R is the usual matrix of 1s and 0s for the row as a categorical variable, and C is the same for the column, our model matrix is X = R − C. This gives us E(y) = Xα. After omitting one of the columns of X due to collinearity, we can then produce the normal least squares estimate α = (X′X)−1X′y. αi −αj = (αi +c)−(αj +c), so {α1, α2, α3, α4, α5} and {α1 +c, α2 +c, α3 +c, α4 +c, α5 +c} are equivalent parameters, demonstrating that one parameter needs to be omitted.

When we preform this procedure on the data, we get these predictions for the α by electrolyte

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Zn | *Fe* | *Pb* | *Cu* |
| O | -0.3908 | -0.8408 | -0.9138 | -1.2426 |
| A | -0.3962 | -0.8690 | -0.9624 | -1.3284 |
| K | -0.4398 | -0.8850 | -0.9566 | -1.3106 |

Table Alpha for the Three Electrolytes

## 3.3 Model Fittting

These two models are one just including the α, and one where the α are interacted with electrolyte. Running these two models and comparing them with ANOVA, we get:

By the interaction, we don’t see quite enough of an improvement to push past the .05 p-value threshold, though it is a borderline case.

## 3.4 Transformation

1. Problem 4

Our data set only has factor covariates, so no monotone transformation will have an effect on the predictions, and non-monotone transformation wouldn’t make sense. If there was evidence of a lack of normality or non-constant variance, a transformation might improve our statistical inference, but checking various residual plots leaves little indication that this is necessary.