601 Final Project

Zhuoyan Xu
zxu444@wisc.edu

University of Wisconsin-Madison

1 Problem 1

1.1 Summary

In this part, we analyzed how the foods calories can be affected by the foods other attributes and whether this attributes can predict calories. Through some analysis on the characteristics of data, I dertermine to use multiple linear regression to achieve our goal.

First, I analyze the variables and collinearity among them, and naively select 9 variables through anova analysis. Then, I check the model assumption and outliers, and applied some remedies on it. Next, I operate the variables selection, and got the best subset of variables if we only pick one variable in one category of attributes. And applied a model validation to check the precision of model. Finally, I use new data to do the prediction and got prediction interval.

1.2 Introduction

The "common household food" dataset include 21 attributes and 961 observations. In this part, we are interested in how the food's calories affected by the food's other attributes. According to the requirements in the assignment, we treat calories as response variable and other attributes as predictors(we may not use all of them) and construct a regression model.

1.3 Part a : Variables and Multicollinearity

The first thing we came up with is which model(linear regression model or generalized linear models) we choose for regression, which mainly depend on the distribution and characteristics of response variables. Since calories is neither a binary data nor a count data, it's continuous and is a physical condition attributes in a proper range, which do not need a link function applied on it. Therefore, we can treat it as normal distribution(which will be carefully checked in the following steps) and apply a multiple linear regression model on it. The multiple linear regression has a couple of coefficients, as we can see below:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_{p-1} X_{i,p-1} + \epsilon_i, \quad for \quad i = 1, 2, \ldots, n,$$

- Y_i is the ith observation of the response variable.
- X_{ik} is the ith observation of the kth explanatory variable for $k = 1, \ldots, p-1$.
- ϵ_i is the ith random error term. The random errors follow a normal distribution with mean zero and variance σ^2 and are independent of each other. We'll focus in this part in the following subsections.

Then, we need to decide how many variables –or more specifically, how many predictors— we should contain in the model and what they are. The ideal scenario is all the predictors are uncorrelated. Each coefficient can be estimated and tested separately. Interpretation such as a unit change in $X_j, j = 1, \ldots, p-1$ is associated with a β_j average change in Y, while holding all other predictors fixed. But if multicollinearity amongst predictors, it is

more difcult to interpret coefficient as the effect of predictors on Y, because the other other predictors cannot be held constant. Also, $X^T X$ is ill-conditioned or rank-decient, which increased calculation complexity.

1.3.1 visualization

The first way to detect multicollinearity is to use scatter plot to visualize them, I want to detect the correlation between variables in each of categories(Fats, Vitamins, Minerals). These categories have more than one variables, which brings in scatterplot matrix to show each pair of variables' correlation. Since the scatter plot and smooth line of each pair of variables are the same in the upper and lower panels, I put the correlation coefficient in the lower panels to see it more intuitively. As we can see:

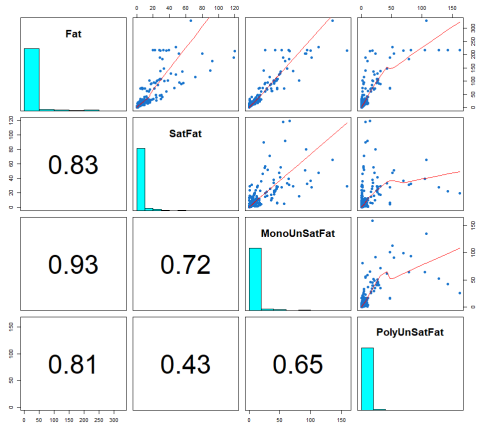

Figure 1: correlation among Fat variables

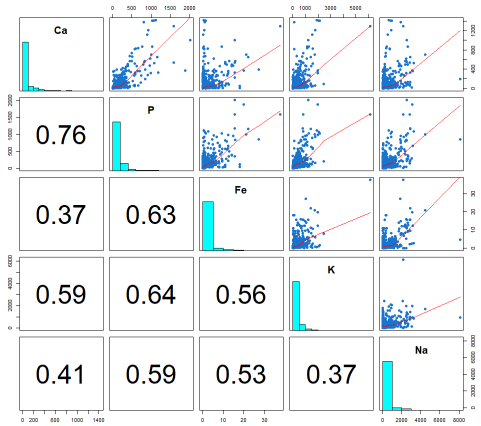

Figure 2: correlation among Mine variables

As we can see in figure 1 figure and 2, which show the collinearity of the variables among Fat category and Mine category. We can conclude there exist serious multicollinearity among variables.

1.3.2 Variance Inflation Factor

Visualization is a good technique to detect collinearity. But if we want to get more precise result we need to do further, that brings in variance inflation factor(VIF). If the mean VIF values of $VIF_k(k = 1, \ldots, p-1)$ is considerably greater than 1, there may be serious multicollinearity problems. If the largest VIF value among $VIF_k(k = 1, \ldots, p-1)$ is larger than 10, multicollinearity may have a large impact on the inference.

When I construct a naive model that contain all the predictors, I calculate the mean VIF is 269.4, the maximum of VIF is over 3000. Thus we can safely conclude there exist serious multicollinearity among variables.

We expect to use variables selection methods to address this problem.For each variable, I delete it in the full model and treat it as reduced model, then I test this variable's significance

by comparing reduced model with the previous full model, the variable who shows significance can stay in the model.

After computation, I got nine variables: weight, protein, Fat, Carbohydrates, Calcium , Phosphorus, Iron, Potassium, Vitamin B1. I let model 1 denotes this model.

1.4 Part b : Model Assumption and Outliers

Multiple linear regression model has several assumptions, we need to check whether these assumptions can be met and make some remedial measures if the assumptions were violated:

- Linearity: A straight line relationship between the response variable Y and the explanatory variable X:

$$E(Y_i|X_i = x_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_{p-1} X_{i,p-1} + \epsilon_i, \quad for \quad i = 1, 2, \ldots, n.$$

- Equal variance: $Var(Y_i|X_i = x_i) = \sigma^2$.
- Independence (conditional on $x_i, x_{i'}$): $\text{Cov}(y_i, y_{i'}) = 0$ for $i \neq i'$.
- Normal distribution:$\epsilon_i \sim N(0, \sigma^2)$.

Where Y_i is i-th observation's response, $X_{i,j}$ is i-th observation's j-th attribute, β_j is j-th parameter, ϵ_i is i-th error term. A very effect way to check these assumption is graphical approach, which is subjective but informative. Since departures from model assumptions can be difcult to detect directly from X and Y. Thus we consider residual plots. Since this model has several predictors, it is reasonable to plot residuals(e_i) against fitted values($\hat{y}_i$). Since raw residuals may show some trend related to predictors, I draw the standardized residuals.

1.4.1 Nonlinearity and Non-equal variance

As we can see in figure 3, the residuals are not showed randomly, its trend and density have some difference upon different fitted values. Thus we can conclude that this model may not meet the linearity.

And also, since the points are not randomly distributed at the both sides on zero, and the range of points is not stable, showing a erratic pattern. Thus the assumption of equal variance cannot be met.

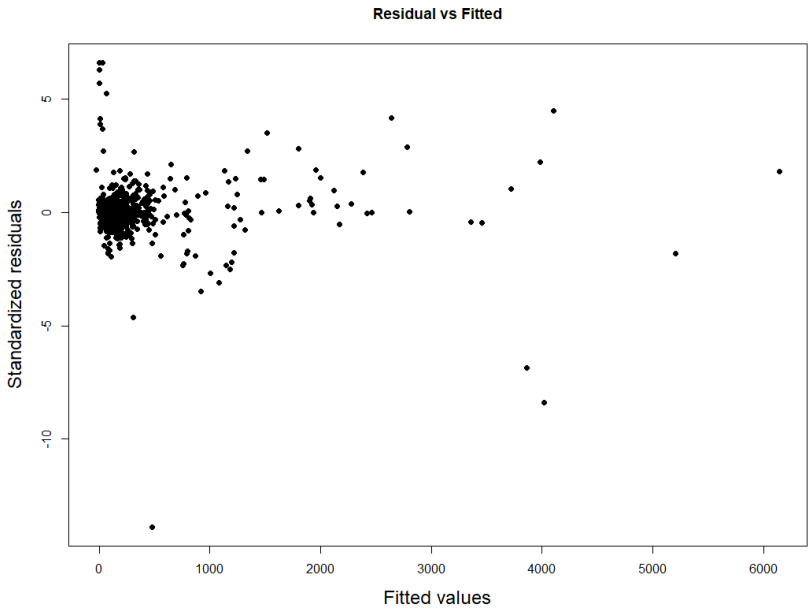

Figure 3: res vs fitted

1.4.2 Nonnormality

If the normality assumption is met, the standardized residuals is asymptotically normal distribution:

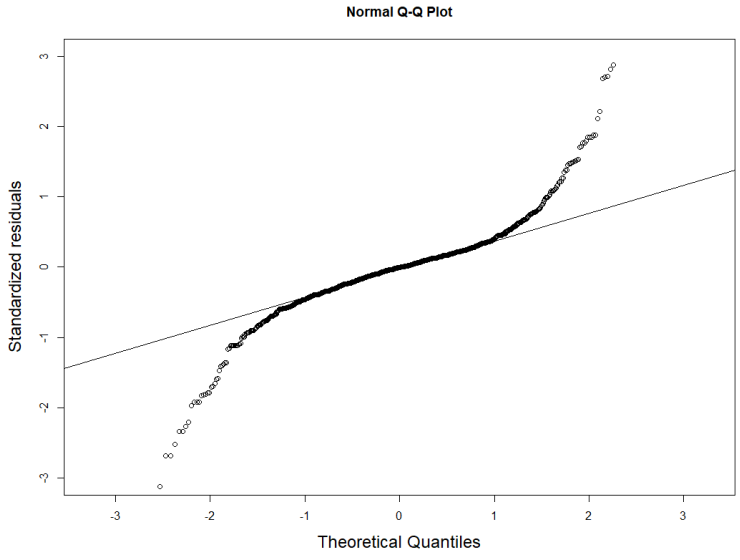

Figure 4: standardized res

As we can see in figure 4, the standardized residuals are asymptotically normal, but it

has a heavy tail, thus we may need some remedies under this situation.

1.4.3 Nonindependence of Error Terms

To check the independence of error term, I plot the residual over index to see whether it is random:

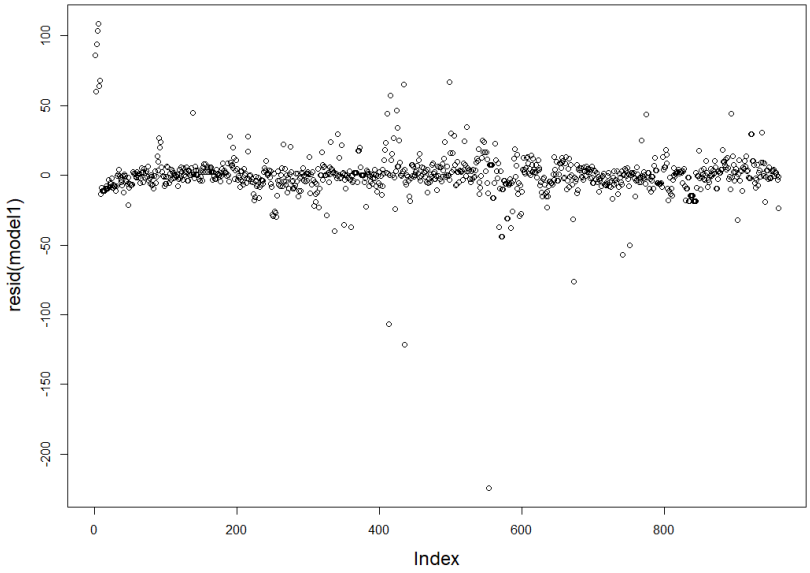

Figure 5: raw residuals

As we can see in 5, the points are randomly distributed and shows no trend upon index. Thus we can say the independence assumption is fairly met according to graphical technique.

1.4.4 Outliers

As we know, if the one observation's studentized residuals are large, what observation may be an outlier. Consider the Bonferroni correction, Under H_0 : If observation i is not an outlier in Y, then the studentized residual (let t_i denotes i-th observation's residual) is t distribution. The decision rule is to reject H0 if $|t_i| > t_{n-p-1}, 1 - \alpha/2n$. by the Bonferroni adjustment for n multiple comparisons, where $t_{n-p-1}, 1 - \alpha/2n$ is higher $\alpha/2n$-th quantile of t distribution with n-p-1 degrees of freedom. If the absolute value of studentized residuals larger than the critucal value(which I draw as the dashed line on the figure) in T test, it can be treated as a outlier:

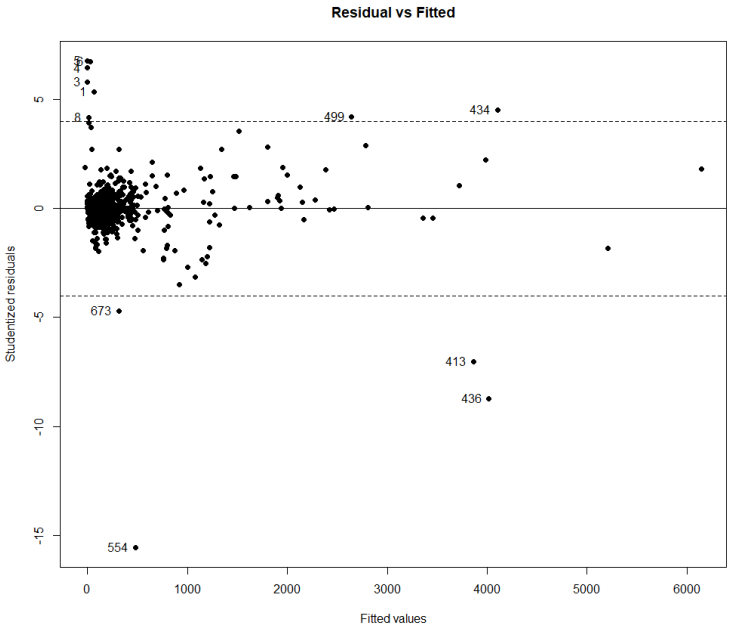

Figure 6: Potential outliers

As we can see in figure 6, the points fall out the dashed line are potential outliers. Then I calculate every observation's value of the DFFITS and Cook's Distance to find influential observations:

- According to the general criterion for DFFITS: for large dataset, the observation whose DFFIRTS value is higher than $2/\sqrt{p/n}$ can be considered to be influential points.
- According to A general rule of thumb for Cook's distance : If observation's Cook's distance is larger than 1, such observation can be treated as possibly inuential.

According to these two criterion, I identify some potential high influential observations. Next, I remove the influential points from the potential outliers. Finally I got one outlier, 8-th observation: WINE; TABLE; WHITE 3.5 F OZ.

1.4.5 Box-Cox transformation

After removing the outlier in the data, I applied a Box-Cox transformation on the data. However, after computation, I got $\lambda = 1$, which means we can get best result among all the conditions if we didn't do transformation. But the problem on the assumption part still exists.

Considering WLS can address the problem of heteroscedasticity, I applied it but the result is not ideal. Therefore, we may need to concentrate on the variable selection part to improve this condition.

1.5 Part c : Variables selection

Since the collinearity mainly among the variables that in the same groups, and referring to the goal "analyze various aspect that may have effect on calories", I decided to select one variable in each group to include in the model.

1.5.1 Stepwise selection

First I considered stepwise selection procedure, which combined forward selection and backward elimination. Considering AIC may result in overfitting especially in large sample, I chose BIC as criterion. Then I got the variables:Fat, Protein, Carbohydrates, Calcium, Phosphorus, Iron, Potassium ,Thiamin and weights.

Then I applied a backwards elimination on a full model, getting the same result.

This result contains 4 variables in Minarals category, and no variables in Cholesterol category, thus this result needs improvement.

1.5.2 Consider all the subsets and choose the best one

In this part, I use the function in "leaps" R package to calculate all the models that contains 7 variables(according to the previous claim), and analyze which is best based on various criterion $(R^2, SSE, R_a^2, C_p, BIC)$.

Then I got the best model contains: Fat, Protein, Carbohydrates, Iron, Potassium ,Thiamin and weights.

This result contains 2 variables in Minarals category, and no variables in Cholesterol category. I tried calculate all the models that contains 8 variables, but that model only add variable Phosphorus comparing to the best model contains 7 variables, and Phosphorus is also in the Mineral category.

1.5.3 Combination

Combine two procedure's result, I add Cholesterol to the model, and choose Potassium over Iron to represent the Mineral category(according to t-test and F-test in anova analysis).

Thus I got my final model which contains: Fat, Protein, Carbohydrates, Cholesterol, Potassium ,Thiamin and weights.

The final model is :

$$\begin{aligned} Y_i = \beta_0 &+ \beta_1 Fat + \beta_2 Protein + \beta_3 Caborhydrates + \beta_4 Cholesterol + \beta_5 Pottasium \\ &+ \beta_6 Thiamin + \beta_7 Weights \end{aligned}$$

1.5.4 Model validation

To test this model's prediction accuracy, I use k-fold cross validation and calculate MSPE, in this situation I let k equals to 10. Then calculate the MSPE:

$$MSPE = \frac{\sum_{i=1}^{n*}(Y_i - \hat{Y}_i)^2}{n*}$$

where n is the sample size of the validation data set. Y_i is the ith observed response in the validation data set. $\hat{Y}_i$ is the ith predicted response in the validation data set. Then I got 10 MSPEs.

The $MSE = SSE/(n-p)$ in this model is 292, there are 3 MSPEs among 10 MSPEs is larger than MSE, the mean of MSPEs is 310.47, which is approximately to MSE. Therefore, we can safely conclude that this model has relatively well prediction accuracy.

1.6 Part d : Prediction

When we have a new observation with predictor X_h according to the requirements in assignment, we can calculate:

$$\hat{Y_h} = X_h\hat{\beta} = 122.0$$

Since the The estimated prediction error variance is

$$\hat{\sigma}_{pred} = \hat{\sigma}\sqrt{1 + X_h(X^TX)^1X_h}$$

Thus we have

$$\frac{\hat{Y_h} - Y_h}{\hat{\sigma}_{pred}} \sim t_{(n-p)}$$

The (1-α) condence interval for μ_h is$\hat{Y_h} \pm t_{(n-p),\alpha/2}\hat{\sigma}_{pred}$

I set confidence level to be 95%. After computation, we have $Y_h \in$ [88.68, 155.30].

1.7 Limitations and Remedies

This model has some limitation since the assumptions are not perfectly met, and still need to be improved in the future analysis:

- According to the assumptions, since the linearity and non-equal variance and normality assumption are not perfectly met, and transformation and WLS regression didn't get very good result, we may need some link function applied on response variable, or considering loess some other non-parametric regression.
- As for variables selection part, choose one variable in one category seems not reliable, we can consider PCA or multidimensional scaling to reduce the dimension meanwhile contain as much information as we can. We can also apply LASSO or ridge regression to achieve bias-variance trade off and reduce the dimensionality.

1.8 Conclusion

In this part, we analyzed how the foods other attributes affect the foods calories and whether can predict calories use these attributes.

First, I detect there is collinearity among the variables and select some variables to construct the model.

Then I check the model assumptions and outliers, I got the 8-th observations outlier, and I removed it.

Next, I use stepwise selection and R_a^2, C_p, BIC *etc.* criterion to got the best subsets if we select one variable in one category, which contains : Fat, Protein, Carbohydrates, Iron, Potassium ,Thiamin and weights. And I use k-folds cross validation to prove the reliability of my model.

Finally, I use the new data to do prediction and got prediction interval. Although this model can get relatively precise conclusion, there are some limitations of this model and need to be improved in the future.

2 Problem 2

2.1 Summary

In this section, we are interested in whether the evaluating of EFRs can predict the audibility of speech.

First, I measure the accuracy that the audibility being predicted by EFRs,and whether this accuracy shows some difference between carriers or frequency groups. I construct a random effect model with accuracy as response and some other variables(such as carriers and participants) as predictors.

Then, I evaluate the performance of two test(F-test and Rayleigh test) for detecting EFRs and choose the more precise one to support the following analysis.

Next, I analyze the connection between SL and EFRs and calculate the minimum SL for EFRto detect a response, and analyze whether it vary by carrier and frequency group.

Finally, I analyze some limatations of my methods and remedies I can operate in the future.

2.2 Introduction

According to the background, we are willing to detect whether EFR has connection with participants' audibility. The experiment was operated on several participants, each participant was tested with different kinds of carriers and different levels of pressure(SPL). The audibility are quantified by the sensation level(SL), which has been got from each experiment's SPL and that participant's known threshold to that sound. Consider each participant was test in several experiments, the detectability within one patient of different SPL might be correlated, thus the result for each observation is not independent, that brings in two procedures : anova model based on categorical variables and random effect model(anova model can also refer to fixed effect model).

According to the condition, each participant have been treated with several carriers and different levels of SPL. Considering anova model, we treat different participant as a factor, which means we believe that the differences among participants are fixed and stable. However, each participant's physical condition is dynamic and difficult to control, it's unreliable to treat it as steady effect. Moreover, if we use anova model, we need to set participant, SPL and carrier as factors, including some significant interactions, that brings in dozens of variables to estimate, which will increase estimate error drastically.

On the other hand, random effect model set carriers as fixed effects(or factors), and set participant and SPL as random effect, which isa error term with expectation zero and variance a constant, which is consistent to the idea that body's physical condition is dynamic and cannot be depict as fixed. This model also consider the dependence of result within one participant of different carriers and SPL.

Therefore, we construct random effect model to achieve our estimation.

2.3 Part a: The estimation accuracy

In this part, we want to measure the accuracy that the audibility being predicted by EFRs, and whether this accuracy shows some difference between carriers or frequency groups.

Consider the detection of EFRs, there are two methods:F-test and Rayleigh test. Two methods have different accuracy. I transfer the p value of test and the SL into binary outcome("P" denotes detectable, "N" denotes undetectable) and compare them. In total experiments, the accuracy of Rayleigh test is 71.73%, the sensitivity and specificity is 67.86% and 91.07%; the accuracy of F-test is 68.75%, the sensitivity and specificity is 63.39% and 95.54%.

At each SPL from the level 25, 35, 50, 65dB, the accuracy of two tests are shown below. As we can see in table 1

	25 dB	**35 dB**	**50 dB**	**65 dB**
Rayleigh	63.10%	46.43%	78.0%	99.4%
F-test	65.48%	41.07%	73.81%	94.64%

Table 1: accuracy

As we can see, the detection's accuracy increases as SPL increases.

Random effect model

We want to detect whether accuracy differ between carriers or frequency groups, which requires us to combine the SL and the test's result into one binary response variables. We have two kinds of test of EFRs, in case for uncertainty, we set accuracy through this criterion: If result of SL and both tests' results are consistent, return to 1, else return to 0. Since response is binary, and can be treat as binomial distribution, I use logit function as link function on the expectation of accuracy.

As I mentioned before, we can use random effect model to achieve our goal. I set carriers as fixed effects, and SPL and participants as random effect, as we can see below:

$$\begin{aligned} y_{ij} &\sim Ber(1, \pi_{ij}) \\ logit(\pi_{ij}) &= \mu + \beta_i + \alpha_j + e_{k[j]} \end{aligned} \tag{1}$$

where y denotes the each test's binary outcome of accuracy, π denotes the each test's true accuracy(range from 0 to 1). $\beta_i = 1, \ldots, 8$ indicates i-th carrier's effect. In generalized random effect model, the residual is excluded by the formula.

There are Two levels of variation:

- At the participant level: $\alpha_j \sim iid\ N(0, \sigma^2)$ are random effects for the participants, j = $1, \ldots, 21$.
- At the SPL level: $e_k \sim iid\ N(0, \sigma_a^2)$ are random effects for the SPL, k = 1,2,3,4.

There are also fixed, unknown parameters:

- μ : overall mean test's accuracy of all experiments.
- σ^2 : variance of accuracy among different participants.
- σ_a^2 : variance of accuracy among different SPL.

After computation, the fixed effects are shown below:

intercept	**/a/F2**	**/i/F1**	**/i/F2**	**/u/F1**	**/u/F2**	**s**	**sh**
-0.20	1.02	1.02	1.60	0.68	0.68	3.13	2.01

Table 2: accuracy upon carrier

As we can see in table 6, the accuracy does differ between carriers, which indicates people's EFR and audibility's connection differ from different kinds of speech stimulus. When carrier is 's', it has the highest accuracy.

Then we still wander do these differences also exist among frequency groups, I construct the same model, get the result shown below:

intercept	**low**	**mid**
2.26	-0.19	-1.39

Table 3: accuracy upon frequency

As we can see in table 3, the accuracy does differ between frequencies, which indicates people's EFR and audibility's connection differ from different frequencies of speech stimulus. When carrier has high frequency, it has highest accuracy.

2.4 Part b : F-test and Rayleigh test

In precious part, accorinf to 1, we got the when SPL is 35,50,65 dB, Rayleigh test has higher accuracy than F-test, at SPL 23 dB, F-test has higher accuracy. We also calculate the sensitivity and specificity of two test. In this part, we can calculate its precision and recall to compare two test's performance, the confusion matrix of two test are shown below:

	Predict P	**Predict N**
Actual P	380	180
Actual N	10	102

Table 4: cm for Rayleigh test

	Predict P	Predict N
Actual P	355	205
Actual N	5	107

Table 5: cm for F-test test

As we can see in table 4, the precision of Rayleigh test is 97.4%, the recall of Rayleigh test is 67.9%.

As we can see in table 5, the precision of F-test is 98.61%, the recall of Rayleigh test is 63.39%.

The Rayleigh test has a better performance on recall, and have little disadvantage on precision.

We are also interested in two test's performance in different frequency, as we can see below:

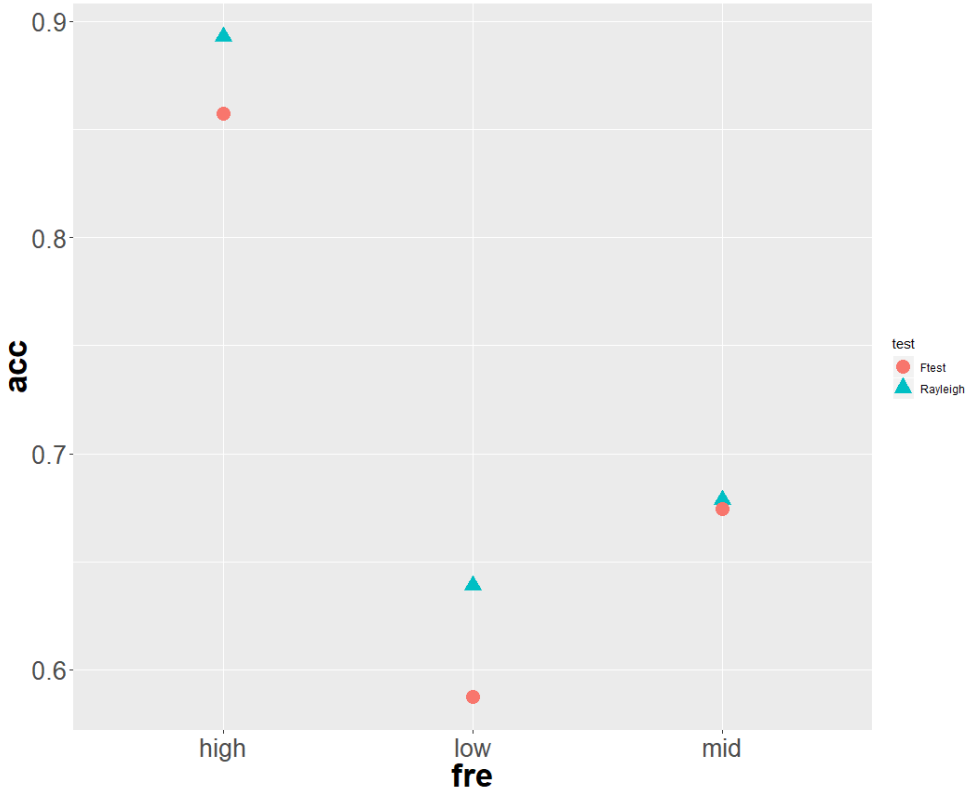

Figure 7: accuracy

As we can see in figure 7, Rayleigh test has better performance over F-test under three frequencies of speech stimuli.

2.5 Part c : SL and EFR

Audibility for each stimulus and each level was quantied by sensation level expressed as dB SL, comparing SPL vs. the participants known audible threshold for that sound we can get every experiment's SL. But when stimulus levels get quieter, we may not be able to detect EFRs at every quite level. Sometimes we need to increase the SPL in order for the EFR to detect a response, but each participent has different threshold at different sounds, we need to determine the threshold for every participant. Considering SL is quantified by SPL and threshold. It's essential for us to determine the minimum SL for the EFR to detect a response.

Thus in this part, we need to estimate critical value of SL at every that EFR can detect a response for every participant. I still use random effect model, but since SL is quantified by SPL, I drop off the SPL in the model, thus our model contains carriers as fixed effect and participant as random effect, and EFR detection(binary outcome) as response variable. Consider the comparison of two test in previous part. Rayleigh has higher accuracy at different levels. We use Rayleigh test result as response variables. The model becomes:

$$\begin{aligned} z_{ij} &\sim Ber(1, \mu_{ij}) \\ logit(\mu_{ij}) &= \mu + \alpha_j \end{aligned} \tag{2}$$

where z denotes whether detect EFR(binary outcome got from Rayleigh test), μ denotes the probability of each test detecte the EFR .

There are only one level of variation, at the participant level: $\alpha_j \sim iid\ N(0, \sigma^2)$ are random effects for the participants, j = $1, \ldots, 21$.

There are also fixed, unknown parameters:

- μ : overall mean test's accuracy of all experiments.
- σ^2 : variance of accuracy among different participants.

Since we predict the probability of whether EFR can be detected, I set 0.5 as threshold, if $\mu > 0.5$, it recorded as detected, else undetected. We need to specify the minimum SL that for all participant can detect EFR, since the critical value of different participant is different, we take the maximum value of it.

After computation, we got the SL is 22.58. It is the minimum SL needed in order for the EFR to detect a response.

Minimum SL in differnt carriers

Since each participent has different threshold at different sounds, we want test whether minimun SL can change if carriers changes. Thus we need to add carrier as fixed effect into model. As we can see:

$$\begin{aligned} z_{ij} &\sim Ber(1, \mu_{ij}) \\ logit(\mu_{ij}) &= \mu + \beta_i + \alpha_j \end{aligned} \tag{3}$$

where z denotes whether detect EFR(binary outcome got from Rayleigh test), μ denotes the probability of each test detecte the EFR . $\beta_i = 1, \ldots, 8$ indicates i-th carrier's effect.

There are only one level of variation, at the participant level: $\alpha_j \sim iid\ N(0, \sigma^2)$ are random effects for the participants, j = $1, \ldots, 21$.

There are also fixed, unknown parameters:

- μ : overall mean test's accuracy of all experiments.
- σ^2 : variance of accuracy among different participants.

I calculate different crotical value SL for different participants at different carriers, and for each carrier, I calculate the minimum SL that all participant can detect EFR, as we can see below:

/a/F1	**/a/F2**	**/i/F1**	**/i/F2**	**/u/F1**	**/u/F2**	**s**	**sh**
32.09	24.20	27.05	20.70	26.73	29.42	10.16	15.01

Table 6: accuracy upon carrier

As we can see in table 6, the minimum SL does vary by carrier, F1 and F2 carriers requires higher SL, thus we can conclude the lower frequency carriers need higher SL to detect EFRs.

2.6 Part d : Limitations and Remedies

This random effect model has some limitations, it's based on some relatively strong assumptions, which need to be checked in future improvements. The model treat the participants and SPL as random effects and carriers as fixed effects, which brings following shortcomings and need to be fixed with some remedies.

1. According to the assumption, the participant's random effect are normal distribution with a constant variance. Although we can treat effects as asymptotic distribution due to large sample, it's unreliable to think participants has the same constant variance. Its variance may based on other effects such as carriers and SPL, thus it's more precise to operate a WLS as remedy.

2. According to the assumption, the participant's random effect are normal distribution with a zero expectation and independent with other effects such as SPL. Since the body's physical condition is dynamic, the fluctuation of threshold for the speech sound may have some trend based on the order of sound pressure level(SPL). For example, one participant threshold for SPL may increases after listening to a higher level SPL sound, it might be harder to recognize a 20dB SPL sound after hearing a 60 dB SPL sound. Thus it's more reasonable to consider the random effect from participant is of regularity based on other effect and find this rule instead of regard them as a random disturbance.

3. According to the assumption, the SPL's random effects are identical independent distributed, which is ill-considered. Since SPL is a sequential level measured in units of decibels, its effects may have some self-correlation, which is ignored in this model. As

for remedies, we may need to construct models with correlated error term, it's a better way to import some methods related to time series, or use some iteration methods to eliminate the impact caused by the self-correlation of the error term.

2.7 Conclusion

In this section, we used random effect model to detect if EFRs can predict the audibility of speech.

1. Trough the estimation of accuracy, we concluded that detections accuracy increases as SPL increases, and when carrier has high frequency, the detection has highest accuracy.

2. Through the comparison of the two test of evaluate whether there is EFRs, we concluded Rayleigh test has better performance over F-test under three frequencies of speech stimuli, thus we used Rayleigh test for major detection of EFRs.

3. Through the connection between SL and EFRs, we calculated the minimum SL for EFR to detect a response, and proved it vary by carrier and frequency group.

4. Although this model can get relatively precise conclusion, it still have some limitations based on its assumptions and need to be improved in the future.

3 Problem 3

3.1 Part a

3.1.1 Conservative matrix

Due to the property of conservative matrix, let $\forall a \neq b \in \{1, 2, \ldots, n\}$, we have

$$V(i_n, i_m) + V(i_m, i_n) = 0$$

thus we have all the matrix V satisfy this equation is skew-symmetric matrix. Thus each conservative matrix is skew-symmetric.

For $\forall a, b \in \mathbb{R}$, and let $\mathbb{V}$ denote the set of conservative matrix, thus $\forall U, V \in \mathbb{V}$, we have

$$(a + b)U \in \mathbb{V} \tag{4}$$
$$a(U + V) \in \mathbb{V} \tag{5}$$

Thus we have the set of conservative matrix is a vector space, and is closed under vector space operations.

This is a skew-symmetric matrix but is not conservative:

$$A = \begin{bmatrix} 0 & 1 & -1 \\ -1 & 0 & 1 \\ 1 & -1 & 0 \end{bmatrix} \tag{6}$$

3.1.2 Additive matrix

Due to the property of additive matrix, we have:for$i_0, i_1, \ldots, i_{n-1}, i_n = i_0$, we have:

$$V(i_0, i_1) + V(i_1, i_2) + \ldots + V(i_{n-1}, i_0) = \alpha_0 - \alpha_1 + \alpha_1 - \alpha_2 + \ldots + \alpha_{n-1} - \alpha_0 = 0 \qquad (7)$$

Thus we have every additive matrix is conservative.

Since every conservative matrix is skew-symmetric, let$k \times k$ V denote any conservative matrix,thus V is also skew-symmetric. We have $\forall i \neq j \in \{1, 2, \ldots, k\}, V(i, j) = -V(j, i)$. There exist α_i, α_j such that $V(i, j) = \alpha_i - \alpha_j$, $V(j, i) = \alpha_j - \alpha_i$.Thus every conservative matrix is additive.

3.1.3 Dimension

Since conservative matrix $V(i, j) = \alpha_i - \alpha_j$ the elements in V is determined by α, thus the dimension of vector space of conservative 6×6 matrices is 6.

A Appendix : The R Code

```
rm(list = ls())
food = read.csv("common_household_food.txt",header = T, row.names =
   NULL)
food = food[-962,]
food1 = food
rownames(food1) = food$Food
food1 = food1[,-1]
###
X = subset(food1,select = -KCal)

qqnorm(food1$KCal)
qqline(food1$KCal)

sum(food1$KCal == 0,na.rm = TRUE)

sum(is.na(food1$KCal))

###--- multicollinearity
panel.hist_line = function(x,...)
{
  usr = par("usr"); on.exit(par(usr))
  par(usr = c(usr[1:2],0,1.5))
  h = hist(x,plot = FALSE)
  breaks = h$breaks;nB = length(breaks)
  y = h$counts;y = y/max(y)
  rect(breaks[-nB],0,breaks[-1],y,col = "cyan")
}
```

```r
panel_cor = function(x,y,digits = 2,prefix = "",cex.cor,...)
{
  usr = par("usr"); on.exit(par(usr))
  par(usr = c(0,1,0,1))
  r = abs(cor(x,y))
  txt = format(c(r,0.123456789),digits = digits)[1]
  txt = paste0(prefix,txt)
  if(missing(cex.cor)) cex.cor = 0.8/strwidth(txt)
  text(0.5,0.5,txt,cex = 5)
}

pairs(food1[,c("Fat","SatFat","MonoUnSatFat","PolyUnSatFat")],upper.
   panel = panel.smooth,lower.panel = panel_cor,
      cex = 1.5,pch = 20,col = "dodgerblue3", bg = "navy␣blue",
      diag.panel = panel.hist_line, cex.labels = 2,font.labels = 2)

pairs(food1[,c("Ca","P","Fe","K","Na")],upper.panel = panel.smooth,
   lower.panel = panel_cor,
      cex = 1.5,pch = 20,col = "dodgerblue3", bg = "navy␣blue",
      diag.panel = panel.hist_line, cex.labels = 2,font.labels = 2)
# pairs(food1[,c("VitaA.IU.","VitaA.RE.","Thiamin","Riboflavin","
   Niacin","VitaC")],upper.panel = panel.smooth,lower.panel = panel_
   cor,
#       cex = 1.5,pch = 20,col = "dodgerblue3", bg = "navy blue",
#       diag.panel = panel.hist_line, cex.labels = 2,font.labels = 2
   )

mod = lm(KCal ~ ., data = food1)
library(car)
vif = vif(mod)
mean(vif)
max(vif)

summary(mod)
anova(mod,test = "Chisq")
summary(aov(mod))
Anova(mod,type = "II")

anova(mod)

###########  b
layout(1)
model1 = lm(KCal~Weight + Protein + Fat + Carb + Ca + P + Fe + K +
   Thiamin,data = food1)
par(cex = 1)
plot(model1$fitted.values,model1$residuals,cex.lab = 1.5)
plot(model1)
```

```
qqnorm(food1$KCal);qqline(food1$KCal)
## fit vs res ##
library(MASS)
plot(model1$fitted.values ,stdres(model1),xlab="Fitted values",
     ylab="Standardized residuals",pch = 16,
     main="Residual vs Fitted",cex.lab = 1.5)
abline(h=0);abline(h=3,lty=2);abline(h=-3,lty=2)
## res QQ ##
qqnorm(stdres(model1),ylab="Standardized residuals",
       ylim=c(-3,3),cex.lab = 1.5);qqline(studres(model1))

plot(resid(model1),cex.lab = 1.5)

## outlier
plot(model1$fitted.values ,studres(model1),xlab="Fitted values",
     ylab="Studentized residuals",pch = 16,
     main="Residual vs Fitted")
abline(h=0);abline(h=4,lty=2);abline(h=-4,lty=2)
text(model1$fitted.values[studres(model1)>4],
     studres(model1)[studres(model1)>4],
     labels = which(studres(model1)>4),pos = 2)
text(model1$fitted.values[studres(model1)<=-4],
     studres(model1)[studres(model1)<=-4],
     labels = which(studres(model1)<=-4),pos = 2)

which(abs(studres(model1))>=4)
outliers = seq(961)[abs(studres(model1))>=4]   ###  potential
   outliers
## DFFITS

lm.reg.dffits = dffits(model1)
plot(lm.reg.dffits, type = "h", ylab = "DFFITS", ylim = c(-3,3))
text(seq(961)[lm.reg.dffits> 2*sqrt(9/961)],lm.reg.dffits[lm.reg.
   dffits> 2*sqrt(9/961)],
     labels = seq(961)[lm.reg.dffits> 2*sqrt(9/961)], cex = 0.8, pos
         = 2)
abline(h = c(-1,-2*sqrt(9/961), 0, 2*sqrt(9/961), 1), lty = 2) #
   specify your own h

seq(961)[abs(lm.reg.dffits)> 2*sqrt(9/961)]

## cook's distance
lm.reg.cooksD = cooks.distance(model1)
plot(lm.reg.cooksD, type = "h", ylab="Cook's Distance",ylim=c(0,2))
text(lm.reg.cooksD, labels = index, cex = 1)
abline(h=qf(0.50,4,15), lty=2) #check whether D_i > f_0.5,p,n-p
```

```
seq(961)[lm.reg.cooksD>qf(0.5,9,952)]


influ = intersect(seq(961)[abs(studres(model1))>=4],union(seq(961)[
   abs(lm.reg.dffits)> 2*sqrt(9/961)],seq(961)[lm.reg.cooksD>qf(0.5,
   9,952)]))
setdiff(outliers,influ)

### transformation
library(MASS)
food2 = food1[-8,]
food2$KCal = food2$KCal+1
bc = boxcox(model1,data = food2[food2$KCal!=0,],lambda = seq(-2,2,0.
   01))
bc$x[which.max(bc$y)]

###-------------c   variables selection
rm(list = ls())
food = read.csv("common_household_food.txt",header = T, row.names =
   NULL)
food = food[-962,]
food1 = food
rownames(food1) = food$Food
food1 = food1[,-1]
food2 = food1[-8,]
X = subset(food2,select = -KCal)

library(leaps)
source("myregsub.R")
N = 1
my = my.regsub(X,y=food2$KCal,nbest=7,nvmax = 8,method="exhaustive")
my1 = my[43:49,]
library(stargazer)

stargazer(my1)


step(mod,direction = "backward",k = log(961))
step(mod,direction = "both",k = log(961))


modelf = lm(KCal ~ Fat + Protein + Carb + Chol + K + Thiamin +
   Weight,data = food2)
summary(modelf)
anova(modelf)
```

```
292

library(caret)
folds = createFolds(seq(960), k = 10, list = TRUE)
MSPE = seq(10)
for(i in 1:10)
{
  model =  lm(KCal ~ Fat + Protein + Carb + Chol + K + Thiamin +
     Weight,data = food2[-folds[[i]],])
  y_hat = predict(model, newdata = as.data.frame(X[folds[[i]],]))
  y_true = food2$KCal[folds[[i]]]
  MSPE[i] = sum((y_hat - y_true)^2)/96
}
MSPE
sum(MSPE>292)
mean(MSPE)

### d prediction
new_X = data.frame(Fat=1.5,
                   Protein = 3,
                   Carb = 26,
                   Chol = 0,
                   Thiamin = 0,
                   K = 95,
                   Weight = 33)
y_pre = predict(modelf,newdata = new_X,interval = "prediction",level
    = 0.95)
y_pre

#
   ###-------------------------------------------------------------------

rm(list = ls())
aud = read.csv("audibility.csv")
aud1 = aud
aud1$Rayleigh[aud$Rayleigh < 0.05] = "P"
aud1$Rayleigh[aud$Rayleigh >= 0.05] = "N"
aud1$Ftest[aud$Ftest < 0.05] = "P"
aud1$Ftest[aud$Ftest >= 0.05] = "N"
aud1$SL[aud$SL < 0] = "N"
aud1$SL[aud$SL >= 0] = "P"

table(aud1$SL,aud1$Rayleigh)
table(aud1$SL,aud1$Ftest)
(107+355)/672
107/112
355/560
```

```
####  accuracy
aud2 = aud1[aud1$SPL == 20,]
#table(aud2$SL,aud2$Rayleigh)
table(aud2$SL,aud2$Ftest)



##  random effect
library(lme4)

## Default REML estimation
aud1$acc = rep(0,672)
aud1$acc[aud1$SL == aud1$Ftest & aud1$SL == aud1$Rayleigh] = 1

fit1 = glmer(acc ~  as.factor(Carrier) + (1|SPL) + (1|Participant),
   family = binomial,data = aud1)
summary(fit1)

aud2 = aud1
aud2$fre = rep(0,672)
aud2$fre[aud1$Carrier == 'a_F1' | aud1$Carrier == 'i_F1' | aud1$
   Carrier == 'u_F1'] = 'low'
aud2$fre[aud1$Carrier == 'a_F2' | aud1$Carrier == 'i_F2' | aud1$
   Carrier == 'u_F2'] = 'mid'
aud2$fre[aud1$Carrier == 's' | aud1$Carrier == 'sh'] = 'high'

fit2 = glmer(acc ~  as.factor(fre) + (1|SPL) + (1|Participant),
   family = binomial,data = aud2)
summary(fit2)

aud2 = aud1[aud1$fre == 'low',]
mean(aud2$Rayleigh == aud2$SL)
aud2 = aud1[aud1$fre == 'mid',]
mean(aud2$Rayleigh == aud2$SL)
aud2 = aud1[aud1$fre == 'high',]
mean(aud2$Rayleigh == aud2$SL)

aud2 = aud1[aud1$fre == 'low',]
mean(aud2$Ftest == aud2$SL)
aud2 = aud1[aud1$fre == 'mid',]
mean(aud2$Ftest == aud2$SL)
aud2 = aud1[aud1$fre == 'high',]
mean(aud2$Ftest == aud2$SL)
```

```
accu = data.frame(acc = c( 0.6388889,0.6785714,0.8928571,0.5873016,0
   .6746032, 0.8571429),
                  test = c(rep("Rayleigh",3),rep('Ftest',3)),
                  fre = rep(c('low','mid','high'),2))
accu
library(ggplot2)
ggplot(accu,aes(x = fre,y = acc,color = test,pch = test)) + geom_
   point(cex = 5) +
          theme(axis.text=element_text(size=20),
          axis.title=element_text(size=26,face="bold"))

###--------------  c
aud2 = aud
aud2$Rayleigh[aud$Rayleigh < 0.05] = "P"
aud2$Rayleigh[aud$Rayleigh >= 0.05] = "N"
aud2$Ftest[aud$Ftest < 0.05] = "P"
aud2$Ftest[aud$Ftest >= 0.05] = "N"

fit3 = glmer(as.factor(Rayleigh) ~  SL + (1|Participant),data = aud2
   ,family = binomial('logit'))
summary(fit3)
coe = coef(fit3)
apply(coe$Participant,2,mean)

ran = ranef(fit3)
ran = ran$Participant
ran = ran$`(Intercept)`

new = data.frame(SL = seq(5,7,by = 0.1),Participant = user[1])
new = data.frame(SL = seq(5,7,by = 0.1),Participant = 'ANH4206')

predict(fit3,new,type = "response")
li = list()

for(i in 1:21)
{
  user = unique(aud2$Participant)
  new = data.frame(SL = seq(-10,50,by = 0.01),Participant = user[i])
  pre = predict(fit3,new,type = "response")
  li[[i]] = new$SL[pre>0.4998 & pre < 0.5002]
  SL = max(unlist(li))
}


## differ in carrier
```

```
fit4 = glmer(as.factor(Rayleigh) ~  SL + (1|Carrier) + (1|
   Participant),data = aud2,family = binomial('logit'))

max(unlist(li))

SL = rep(0,8)
for(j in 1:8){
  cari = unique(aud1$Carrier)

  for(i in 1:21)
  {
    user = unique(aud2$Participant)
    new = data.frame(SL = seq(-10,50,by = 0.01),Carrier = cari[j],
       Participant = user[i])
    pre = predict(fit4,new,type = "response")
    li[[i]] = new$SL[pre>0.4998 & pre < 0.5002]
    SL[j] = max(unlist(li))
  }

}
```