Final Project, STAT 849

Yunyi SHEN

12/13/2019

Calories

Abstract

I analyzed the data set of common household food's calories and nutritional profile. Linear regression and LASSO were used in order to obtain empirical formula to predict calories density (defined as calorie divided by serving weight, in KCal/g) using measurement of nutritional profile (weight of such nutrient per gram of food). I obtained two such formula for different goals, 1) for predicting calories using 6 classes of nutrients. For this goal, best predictors were total fat, carbohydrates, protein, cholesterol, iron and vitamin A in IU unit. 2) for predicting calories using minimal nutrient measurements. For this goal, best predictors were total fat, carbohydrates and protein. Two models achieve similar prediction power which was measured using 10-fold cross validation and mean square prediction errors were 0.0641 (SD=0.012) and 0.0755 (SD=0.012), respectively. An unknown food item's calorie was predicted by two models and results were 50.3KCal (0.95 prediction interval= $(34.1, 66.5)$) and 49.5KCal, respectively.

Background

Calories are essential to human health. How calories are different among different type of food was an foundation for nutritional science. We collected calories and nutritional content (e.g. amount of fat) data for 948 common household foods. Our goal in general was to understand which nutrition's effect on calorie level and produce a model for predicting calories from nutritional profile of the food.

Nutrition can be divided into 6 main classes, naming fat, Protein, Cholesterol,Carbohydrates,Minerals and Vitamins. One of my goal was to find no more than one best representative in each of these 6 classes that together can best predict calories. Also, in some cases especially in engineering, researchers are more willing to obtain a model with least measurements while keeping an acceptable accuracy compare with a model with more measurements. This was the second goal of this analysis, i.e. finding a model using least predictors while keeping an acceptable performance of predicting.

There are various way to address this feature selection problem. For the first goal, as we have small number of predictors, model selection based on Akaike's information criterion (AIC) could be used to select the model. For the second goal, LASSO can offer the sparse solution of this regression problem which fits the goal of having least predictors. As we mainly focus on prediction in this analysis,cross validation was to test the predicting power of both models.

This study proposed two empirical formula for predicting calories from nutritional profile. In total they used 6 and 3 predictors. Two formula could be chosen by users based on their goal.

Material and Method

Preparation Before Analysis

Because of the additive property of calories (in KCal), it is proportional to the weight of the food. To account for variance of weight in different food I first standardized all variables by the serving weight of the food, make the data being calorie density (in KCal/g) and portion of certain nutrient (e.g. mg VC/g). Because physically calorie is additive through out all ingredient, I did not do any transformation on calorie density. All the predictors were Z-standardized (i.e. linear transformed to obtain a 0 mean and 1 standard deviation, see Table 1 for the empirical mean and standard deviation). Because water is the product rather than reactant of major respiration reactions of human (Nelson and Lehninger 2008), together note that water has a large variance inflation factor (VIF) value of 78.85, water was not included as predictor in all the model constructions. I will later work on this reduced and standardized data set for analysis.

Simple Linear Regression and Model Selection for Representitive of 6 classes of Nutritions

In order to analyze the relationship between calorie level and nutritional content, I used linear regression between calorie and nutritional contents. This assumes measurement error on calories were independent identically normal distributed.

My ultimate goal is to obtain models predicting calorie density from nutritional profile. There are two subgoals. The first goal was to predict calorie density from major classes of nutrients: I divided all nutrition into 6 categories, i.e. Fats, Proteins, Carbohydrates, Cholesterol, Vitamin and Minerals (Table.1). I chose at most one predictor from each category and generated in total 1680 models. Models were compared using AIC. The model with least AIC value was chosen for prediction in this subgoal. Variance inflation factor (VIF) of predictors were calculated after conducting any regression to check multicollinearity problem. If maximum VIF is larger than 10, that model was investigated further. One independent sample was used to test the prediction power of the model. Standard model diagnostic measures was visually checked. Outliers were checked by Cook's distance with cutoff 1. I also conducted a 10-fold cross validation for this model to evaluate its predicting power.

LASSO Regression for Least Measurement Prediction

Second subgoal was to predict calorie density from least measurements while keeping a reasonable predicting power. Before conducting any regression analysis, I removed several redundant measurements based on previous studies. Four redundant measurements were removed: `SatFat`, `MonoUnSatFat`, `PolyUnSatFat` because for calorie contribution, these fat are similar thus we can use total fat `Fat` to represent. In food

Table 1: Summary of predictor used and their grouping

	Abbreviation	Mean	SD	Unit	Group
Protein	Protein	7.23	10.1	g	Protein
Total Fat	Fat	12.5	33.1	g	Fat
Saturated Fat	SatFat	3.99	10.7	g	Fat
Monounsaturated Fat	MonoUnSatFat	4.89	14.1	g	Fat
Polyunsaturated Fat	PolyUnSatFat	2.83	11.5	g	Fat
Cholesterol	Chol	32.6	120	mg	Cholesterol
Carbohydrates	Carb	34	78.5	g	Carbohydrates
Calcium	Ca	78.8	165	mg	Minerals
Phosphorus	P	130	205	mg	Minerals
Irons	Fe	1.78	3.14	mg	Minerals
Potassium	K	275	382	mg	Minerals
Sodium	Na	320	626	mg	Minerals
Vitamin A (IU)	VitaA.IU.	1040	3860	IU	Vitamins
Vitamin A (RE)	VitaA.RE.	149	509	RE	Vitamins
Vitamin B1	Thiamin	0.167	0.307	mg	Vitamins
Vitamin B2	Riboflavin	0.199	0.362	mg	Vitamins
Vitamin B3	Niacin	1.91	3.19	mg	Vitamins
Vitamin C	VitaC	11.2	32	mg	Vitamins

energy related literature, total fat were widely used (e.g. Rolls 2000, Rolls and Hammers 1995, Warwick and Schiffman 1992) . I also removed `VA.RE.` because we also used IU to measure VA profile. LASSO was used to obtain a prediction model with least predictors with an acceptable predicting power. I used 10-fold cross validation to obtain the shrinkage parameter λ. For robustness and the goal of least measurements, λ corresponding to the most regularized model (i.e. least measurements needed) such that MSPE is within one standard deviation of the minimum (keep reasonable predicting power) was chosen for prediction. The λ gives the lowest estimated MSPE was used for comparison. This model will also be compared with the least AIC model in terms of MSPE estimated by 10-fold cross validation.

All analysis was done in `R 3.6.1` (R Core Team 2019) and LASSO was done using `R` package `glmnet` (Friedman et al. 2010).

Results

Predicting Food Calories from Class of Nutrients

I ran 1680 models in total and no model has maximum VIF>10 (Fig.1 in appendix for the statistics of maximum VIF values of all models). There was no point with Cook's distance>1 in the model had least AIC (least AIC model later). Standard model diagnostic did not show obvious violation of assumptions in the least AIC model. Table. 2 showed predictor used in least AIC models as well as Lasso model. In the least AIC model, `Fat`, `Protein`,`Chol`, `Carb`,`Fe` and `VitaA.IU` were selected for prediction. 10-fold cross validation shows the MSPE is 0.064, with sd=0.012 implied a good prediction power of the least AIC model.

Explanation of model coefficients (Table.2) should be done with cautions. Because predictors were Z-standardized and thus not on the original scale. Parameters should be understood as calorie density change when such nutrient increased by 1 standard deviation (Table.1), while other nutrients hold constant. In this sense, fat has the largest influence on food calorie, when increased by 1 standard deviation, calorie density will increased by 1.8KCal/g, carbohydrates and protein were the second and third, with increase of 0.95 and 0.35 respectively. Follow that, cholsteral (0.025), iron (-0.031) and VA in IU (0.033) has much smaller coefficients.

Based on my result I proposed this empirical formula to predict calorie density (in KCal/g) from 6 class of nutrients:

$$\begin{aligned}\frac{Calorie}{Weight} &= 2.25 + 1.8(\frac{Fat/Weight - 12.5}{33.1}) + 0.35(\frac{Protein/Weight - 7.23}{10.1}) + 0.95(\frac{Carb/Weight - 34}{78.5}) \\ &+ 0.025(\frac{Chol/Weight - 32.6}{12.0}) - 0.031(\frac{Fe/Weight - 1.78}{3.14}) + 0.033(\frac{VA.IU/Weight - 1040}{3860})\end{aligned} \tag{1}$$

This formula can be used on the original scale of nutrients, however cautious should be paid to the unit of such measurements, see Table.1 for a summary. Test result for the unknown food item was given in Table.3.

Predicting Food Calories from Least Measurement of Nutrients

Fat, protein and carbohydrates were chosen by LASSO out of 14 candidate predictors using the 1se rule. Model performance evaluated by MSPE indicated that this simplest model has only minor performance lose (MSPE(sd)=0.076(0.012), within 1 stand deviation of MSPE compare with the least AIC model (`least AIC` in Table.2, MSPE(sd)=0.064(0.012)) as well as LASSO chosen using the least MSPE rule (`LASSO-min` in Table.2, MSPE(se)=0.065(0.012)). However, number of measurement required decreased by 3 and 2 compare with least AIC and least MSPE LASSO model.

From LASSO result, I proposed a empirical formula for predicting calories from total fat, carbohydrates and protein content:

Table 2: Model Coefficients and 10-fold Cross Validation Mean Squared Prediction Error (MSPE) of least AIC and LASSO models. Predictors Unshown were Unused

	least AIC	LASSO-1se	LASSO-min
intercept(se)	2.25(0.00803)	2.25	2.25
Fat(se)	1.76(0.00825)	1.7	1.74
Protein(se)	0.351(0.00877)	0.296	0.331
Carb(se)	0.945(0.00884)	0.858	0.911
Chol(se)	0.025(0.00869)	-	0.0143
Fe(se)	-0.0308(0.00957)	-	-
VA.IU(se)	0.0328(0.00899)	-	0.00154
MPSE(sd)	0.0641(0.012)	0.0755(0.012)	0.0653(0.012)

Table 3: Point Prediction and 0.95 Prediction intervals of least AIC and LASSO model for new food item

	fit	lwr	upr
least AIC	50.3	34.1	66.5
LASSO-1se	49.5	-	-

$$\frac{Calorie}{Weight} = 2.25+1.70(\frac{Fat/Weight-12.5}{33.1})+0.30(\frac{Protein/Weight-7.23}{10.1})+0.86(\frac{Carb/Weight-34}{78.5}) \quad (2)$$

Point estimation of coefficients of common predictors (fat, cabohydrat and protein) are similar, indicate the consistency of these two models. User could chose between these two based on their goal.

Discussion

Justification of Classification of Nutrients

Despite the previous knowledge about the classification of nutrients, we can still observe the clustering nature of the nutrients by looking at their correlation plot.

Thus it make sense to select one out of the each class for predicting but not for inference on which nutrient has the largest influence on calories for two reasons: first we cannot infer causation from correlation, second if several nutrients have similar source from food, then food data cannot used to evaluate the calorie change if holding one constant and change another.

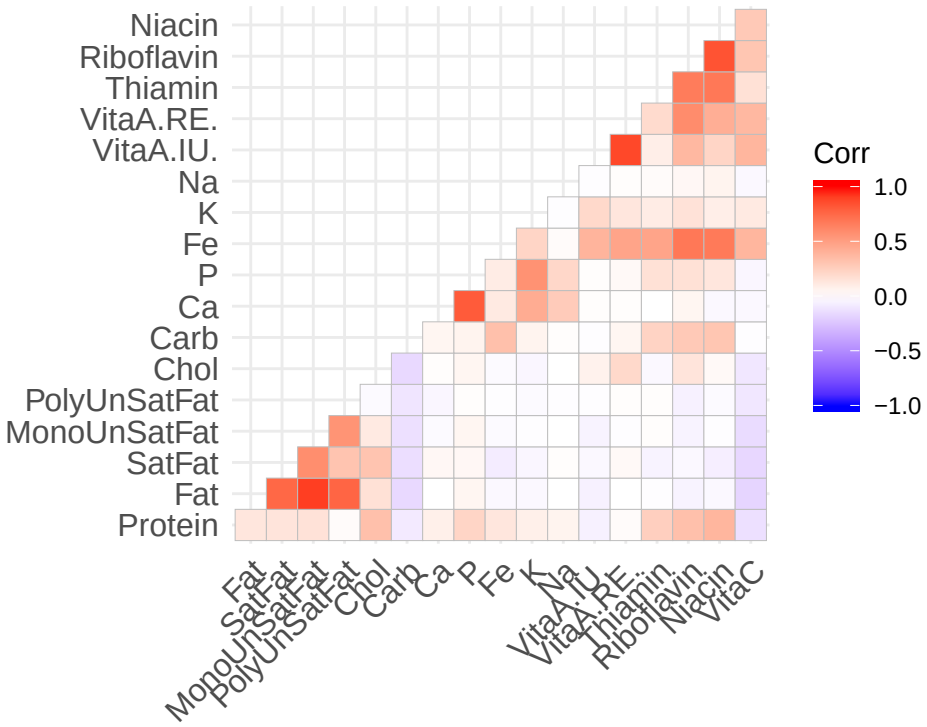

Figure 1: Correlation plot of all candidate predictors

Investigation of Points with Large Residuals

Three points had large residuals in the least AIC model (Fig.2 in Appendix), point 3,4,and 5. They are all Shisky with proof of 80, 86 and 90 respectively. All residuals are positive (2.241,2.45,2.57, while residual standard deviation was estimated as 0.2488). Since they are all ethanol rich drinks, one possible explanation is because we did not measure ethanol level in these drinks. Bomb calorimetry was widely used in measuring food calories (e.g. Acheson, et al. 1980). In this method, people combust the food in a chamber, with water cover around the chamber then measure the temperature change of the water to obtain the energy of combustion. Ethanol can be combusted and contains some amount of energy (with combustion enthalpy -1370.7 kJ/mol = 7.12 KCal/g). From here we can calculate the calories from ethanol, for 80-PROOF, the amount of ethonal can be estimated from the remaining mass after taking out water, which is $42 \times (1 - 0.67) = 13.86g$, the energy from ethanol is given by $13.86 \times 7.12 = 98.7KCal$ and then the residuals for calorie densities are 0.0088,0.0638 and 0.0875 respectively, which is lower than the residual predicted by the model. This may explain the residual of point 3 to 5. Detailed knowledge about the experimental design underlying this data set is necessary for further understanding of these large residuals. Since it was not considered as outliers due to Cook's distance, I will not remove them before obtaining further information about how the experiment was conducted.

Major Calorie Bearing Nutrients

The predictor picked by LASSO agree with the previous knowledge of that fat, carbohydrates and proteins are major energy bearing nutrient (Nelson and Lehninger 2008). Consider point estimation of the coefficient corresponding to other predictors in the least AIC model, they are all smaller than those of fat, carbohydrates and proteins by 3 orders of magnitude. These observations were consistent with the knowledge long been known: fat, carbohydrates and protein explains more of the variance in food calories than other nutrients. Again, explanation of there parameters should be careful, since all predictors were Z-standardized, one unit change correspond to changing by one standard deviation rather than its original unit.

However, since there were multicollineraty (Figure.1), from this analysis we cannot conclude that these three nutrients contributed most of the calories in food items in a causation sense. The empirical formula can be useful for predicting but it may not be a physically correct model to explain the source of calories.

Implication on Calorie Prediction

I proposed two empirical models for users to choose based on their goal and measurement available. If the goal is to predict calories from least measurement, LASSO model can be used, with only 3 measurement needed. If their goal is to predict calories from measurements of classes of nutrients, least AIC model can be used. Least AIC model suggested measure should be taken on total fat, protein, cholesterol carbohydrates, iron and vitamin A in IU unit. Units should be same with Table.1.

End of the Main Report

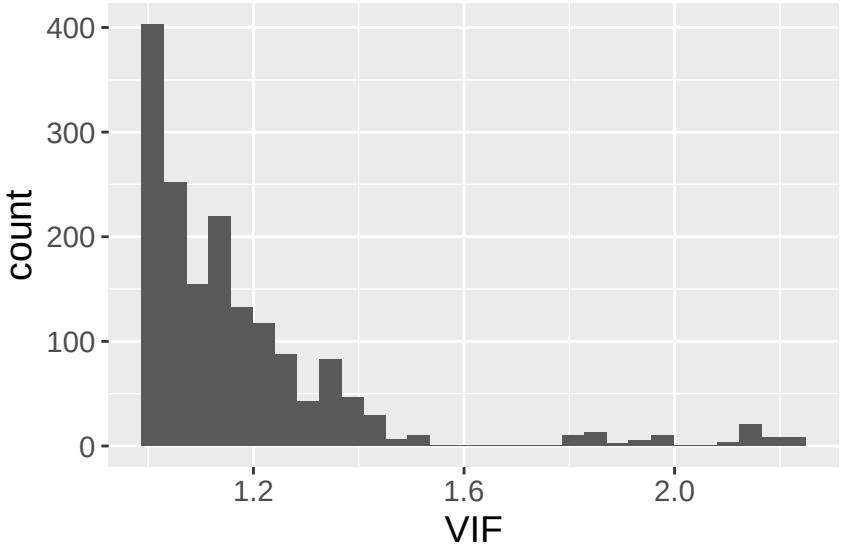

Figure 1: Histogram of maximum VIF values for all ardinary least square linear model fitted

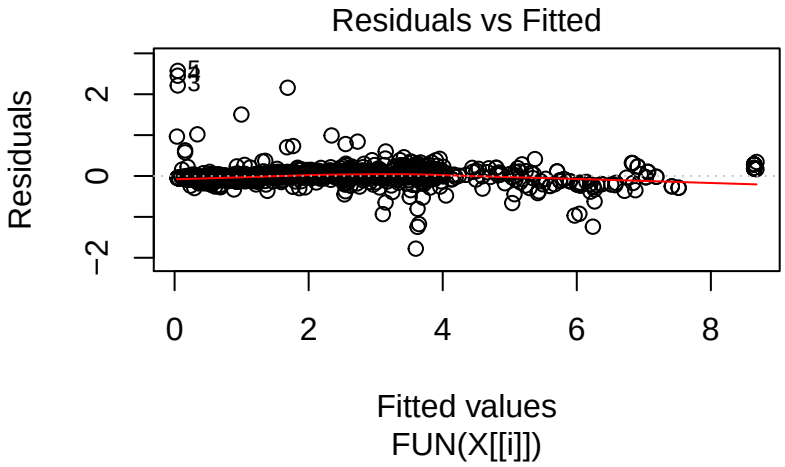

Figure 2: Residual plot for least AIC model

Appendix

Two figures were in the appendix. Figure.1 is the histogram of all model's largest VIF score. Figure.2 is the residual vs fitted plot of the least AIC model.

References

Acheson, K. J., et al. "The measurement of food and energy intake in man—an evaluation of some techniques." The American Journal of Clinical Nutrition 33.5 (1980): 1147-1154.

Friedman, Jerome, Trevor Hastie, and Rob Tibshirani. "Regularization paths for generalized linear models via coordinate descent." Journal of statistical software 33.1 (2010): 1.

Institute of Medicine (US) Standing Committee on the Scientific Evaluation of Dietary Reference Intakes. Dietary reference intakes for thiamin, riboflavin, niacin, vitamin B6, folate, vitamin B12, pantothenic acid, biotin, and choline. National Academies Press (US), 1998.

Nelson, David L., Albert L. Lehninger, and Michael M. Cox. Lehninger principles of biochemistry. Macmillan, 2008.

R. Core Team. "R: A language and environment for statistical computing." (2019) https://www.R-project.org/

Rolls, Barbara J. "The role of energy density in the overconsumption of fat." The Journal of nutrition 130.2 (2000): 268S-271S.

Thompson, June. "Vitamins, minerals and supplements: part two." Community practitioner 78.10 (2005): 366.

Warwick, Zoe S., and Susan S. Schiffman. "Role of dietary fat in calorie intake and weight gain." Neuroscience & Biobehavioral Reviews 16.4 (1992): 585-596.

Audibility

Abstract

Researchers proposed to use envelope following responses (EFRs) to predict whether a speech was audible or not. I compared amplitude (F-test) based and phase based (Rayleigh-test) methods for detecting EFRs with the sensation level (SL) which was a benchmark method for audibility in 8 carriers that was divided into 3 frequency categories. In total 672 trails were conducted during which 21 participant involved and 8 different carries each had 4 different sound pressure levels. I used receiver operating characteristic (ROC) curve and the area under the curve (AUC) to evaluate the performance of EFR in predicting audibility assigned by SL method. I fail to detect differences in the AUC values between amplitude and phase based methods in either overall (p=0.56 DeLong test, H_0: no difference in AUC value) or carrier-specific situations (maximum p=0.093 DeLong test H_0 same as before). Two tests had highest AUC on carrier s (0.89,0.93 for amplitude and phase) and sh (0.84 for both amplitude and phase) and lowest on carrier /a/F1 (0.64 and 0.60 for amplitude and phase). Afterward I set cut off to be 0.05 convert EFR detection to a binary variable, and perform a χ^2 independence test between consistent prediction with SL and carrier/frequency on both F and Rayleigh test results. After correction of p-value for multiple comparison, I still detect dependency between the consistent prediction and carrier (p=2.83e-8,1.83e-7 for F and RayLeigh with H_0:number of consistent prediction is independent with carrier) as well as frequency (p=3.26e-8,2.47e-8, with similar H_0 but for frequency). This result suggested that performance of EFR methods depends on carrier and frequency. To Evaluate this dependency, for each EFR detection method I conducted 10 logistic regression analysis between EFR detection and SL, carrier as well as frequency, in 5 of them included individual as random effect on intercept. Akaike information criterion (AIC) was used to determine the best model for prediction. Best model for two tests had similar structure i.e. included random effect, intercept was same among carriers and slope was different among carriers. 10-fold cross validation were conducted and ROC-AUC were calculated to asses their predicting power on ESR. SL threshold for EFR detection was defined as the value SL to have 0.5 chance of getting an EFR detection. Such thresholds were calculated and their bootstrap 0.95 confidence interval (CI) was obtained. I found all thresholds were greater than 0 and their 0.95 CI did not overlap with 0 suggested that EFR based method were less sensitive than SL method. And in general high frequency groups had lower threshold. Carrier `s` has the lowest threshold of 6.03 (0.95 bootstrap CI: 4.08-8.01) for F-test and 3.36 (0.95 boot strap CI: 1.58-4.67) for Rayleigh test. Carrier /a/F1 had the highest of 33.3 (0.95 bootstrap CI: 26.1-41.5) for F-test and 25.8 (0.95 bootstrap CI: 19.3,33.5) for Rayleigh-test.

Background

Researchers proposed a new method called envelope following responses (EFR) to evaluate audibility of certain speech. EFR was based on electro-encephalogram (EEG) which was a real-time data collection method that flourished recently. EFR, the new EEG based method has a potential to be widely used by researchers

interested in real-time evaluation of audibility. However before using it for further research, validation was needed. To validate this newly proposed model, researcher was interested in compare it with a benchmark method that using sensation level (SL, positive means audible).

Two method for detecting an EFR were used. One compares the amplitude of the EFR signal with noise amplitude using F-test. Another compares the inter-trial consistency of phase using Rayleigh test. Two methods both produced a p-value as predictor of EFR detection. Researchers considered $p<0.05$ to be a detected EFR.

It was also of interest whether there exist any systematic pattern of EFR detectability. Researchers used different frequency and sound pressure for different trail in order to detect such pattern.

To validate this new method, I compared detected EFR and audibility based on benchmark SL method. I evaluated number of consistent predictions between EFR and SL in variance settings. Further, to determine the threshold of SL to detect a EFR, I used regression analysis between EFR detection (at 0.05 level) and their corresponding SL, as well as carriers/frequencies as predictors.

Material and Method

Experiment

In total 21 participant involved in this experiment. Each participant was given 32 speeches with 8 different carriers and each carriers had 4 different sound pressure level (SPL). Benchmark was taken using sensation level (SL) method. Meanwhile, participant's EEGs were recorded for analysis. Two method were used to identify an EFR: compare amplitude with noise using F-test and compare inter-trial phase consistency using Rayleigh-test. Test p-values were recorded.

Comparing F-test and Rayleigh test for Detecting Audibility

In this section, first goal is to compare two test's predicting power on the whole data set. We can view the two tests as two classifiers whose predictors were p-values and our goal in this section was to compare these two classifier's performance using ground truth given by SL method here.

To reach our goal, I did not use any pre-specified cut-off p-values for detection, instead I used receiver operating characteristic (ROC) curve and the area under the curve (AUC) to evaluate their performance. SL was used as ground truth. Speech with $SL > 0$ was considered as audible (positive). AUC was calculated to be a measurement of overall performance for each test. Because two tests were testing the same individual and trial, I used DeLong's test (DeLong et al. 1988) for correlated AUCs to test the null hypothesis that two AUCs are the same (i.e. there is no proformance difference between F and Rayleigh test).

Later, I divided the data set in to 8 sub data sets based on carriers and conducted the same ROC-AUC test for each of them. Due to multiple inferences, I adjust the p-value cutoff for significant to $0.05/9 = 0.00556$.

Evaluating EFR's Prediction on SL based Audibility

In this section, the goal is to check whether a detected EFR can predict a positive SL, or vise versa. I used p<0.05 as cutoff of detected EFR in both F-test and Rayleigh test results. I use SL>0 for benchmark audible. I evaluate the consistent rate as a measure of accuracy, defined as number of consistent predictions (i.e. detect EFR and $SL > 0$, or no detected EFR and $SL < 0$) divided by total number of trails.

Further, to test whether accuracy is different between carriers or frequency groups, I used Pearson's χ^2 test of independency between consistent prediction and carriers/frequency groups respectively. The null hypothesis of these tests were consistent prediction was independent with carriers or frequency groups.

Evaluating Minimum SL Needed for EFR Detection

In this section, the goal is to find the minimum SL needed for EFR to be detected. Using cut off $p < 0.05$ I convert EFR to a binary responses. With the assumptions participant's EFR detection were independent to each other and follow Bernoulli distribution. I constructed 5 mixed effect logistic regressions using individual as random intercept and SL and carrier/frequency as predictors. I constructed 5 fixed effect logistics regressions using SL and carrier/frequency as predictors (this further assumed that different trails on one participant were independent). For each class of logistic regressions (i.e. fixed and random effect), 5 regressions correspond to:

1. Only SL has the main effect;
2. All groups share the intercept, slope differs among different frequency (interaction between SL and frequency, no main effect);
3. All groups share the intercept, slope differs among different carriers (interaction between SL and carrier, no main effect);
4. Slope as well as intercept differs among different frequency (interaction between SL and frequency, with main effect);
5. Slope as well as intercept differs among different carriers (interaction between SL and carrier, with main effect).

In which, slope should be understand as the log detection odds ratio change when SL increased by 1 unit and intercept was the baseline probability of detecting a EFR when SL is 0.

Akaike information creterion (AIC) was used to choose the best model for calculating the minimum SL needed, defined as the SL value that there was a 0.5 chance to have a EFR detection. For mixed effect models, 95% confidence intervals were calculated using bootstrap method offered in `R` package `lme4` (Bates et al. 2015) and `boot` (Canty and Ripley 2019). I took 500 bootstrap samples to construct such confidence interval (CI) using percentiles of bootstrap samples.

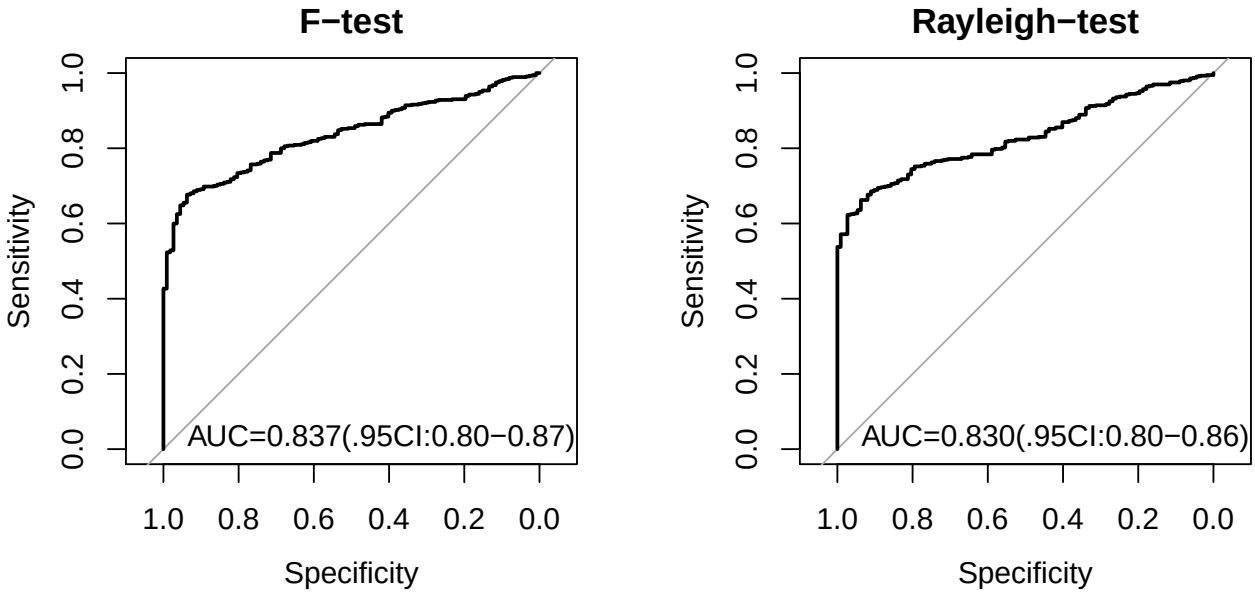

Figure 1: Overall ROC curves for Rayleigh and F test, DeLong test gives a p-value=0.56 with null hypothesis: area under two ROC curves are the same

Bootstrap process will resample the data with replacement and fit the same models for 500 times, given estimation of some quantities that of interest. Then we can take the sample as an approximated sample distribution of such estimates. I took 0.975 and 0.025 percentile to construct the 0.95 CI for thresholds.

Prediction power of this least AIC model was also estimated using ROC-AUC method with 10-fold cross validation. Point estimations (i.e. point estimation of threshold and response curve) were calculated by taking average of all 10 models first on logit scale and transform back to raw scale. Then I predicted the response curve assuming the random effect is 0, i.e. an average patient. I first predict the response on logit scale by 10 models in the cross validation process and take inverse logit to get the response curve. Subset of data was chosen using `R` package `caret` (Kuhn et al. 2019).

Results

Comparing F-test and Rayleigh test

ROC curves for overall sensitivity and specificity of the two tests were shown in Figure.2. DeLong test gave a p-value=0.56 with H_0 : there is no difference between two test's AUCs. Thus we did not have evidence that the overall performance of this two test differs on this particular data set. For each specific carrier, DeLong test gave all p-value$>$ 0.005, we have no evidence that performance of these two tests differs on any of the 8 carriers (Figure. 2). This suggested there may not be performance difference between amplitude (F-test) and phase (RayLeigh-test) based EFR detection.

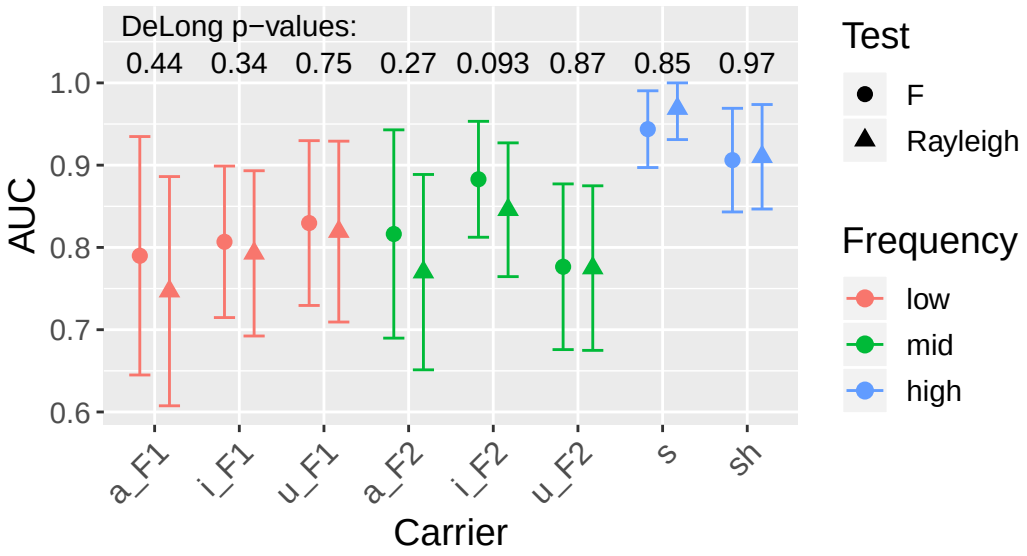

Figure 2: AUC and corresponding DeLong 0.95 CI for each test on each carrier, p-value at the top of each carrier was for DeLong-test with H0: on such carrier, two test had same AUC value

EFR's Prediction Power

F test result was consistent with SL results in 462 out of 672 trails (68.75%), while Rayleigh test was consistent in 482 trails (71.73%) when chose p-value<0.05 as detected EFR. Sensitivity, defined as rate of detecting EFR when such speech has a positive SL, was 63.39% for F-test and 67.85% for Rayleigh test. Specificity, defined as rate that speech had a positive SL when an EFR was detected for such speech, was 98.6% and 97.4% for F-test and Rayleigh test respectively (Table.1, first 3 columns).

The χ^2 independent test between consistency (i.e. whether EFR made the same prediction of audibility with SL) and carrier whose H_0 was that consistency was independent with carrier had p-value 2.83e-8 and 1.83e-7 for F-test and Rayleigh test respectively. The χ^2 independent test between consistency and frequency whose H_0 was that consistency was independent with frequency had p-value 3.26e-8 and 2.47e-8 for F-test and Rayleigh test respectively (See appendix Table. 1 to 4 for the contingency table). Due to multiple comparison, we need to adjust our p-value cut off for significant to $0.05/4 = 0.0125$, however we still have a strong evidence for that accuracy differ between carriers and frequency groups (Table.1, last 2 columns). From AUC of each carrier, we can observe that both F and Rayleigh test were good at high frequencies (carriers s and sh).

Minimum SL needed for EFR

In total 10 models were constructed for each detecting method respectively (F and Rayleigh). The mixed effect model used same intercept among carriers, while slope depended on carriers was selected based on the least AIC rule (Table.2) for both of the two detecting method. ROC-AUC showed a good predicting power of this model (AUC=0.921(SD=0.0224),0.932(SD=0.0393) for F and Rayleigh).

Table 1: Performance summary of two tests using 0.05 as cut off and chisq independent test with H0: overall correct is independent with carrier/frequency

	Overall	Sensitivity	Specificity	p-value Chisq test with carrier	p-value Chisq test with frequency
F	0.688	0.634	0.986	2.83e-08	3.26e-08
Rayleigh	0.717	0.679	0.974	1.83e-07	2.47e-08

Table 2: deltaAIC table for 10 candidate logistic regression models on SL and EFR detection

random effect	intercept	slope	deltaAIC_F	deltaAIC_Rayleigh
individual on intercept	same among carriers	differ among carriers	0	0.000
individual on intercept	differ among carriers	differ among carriers	7.47	0.427
none	same among carriers	differ among carriers	12.7	37.300
none	differ among carriers	differ among carriers	21	39.600
individual on intercept	same among frequencies	differ among frequencies	21.2	17.300
individual on intercept	differ among frequencies	differ among frequencies	21.4	14.800
none	same among frequencies	differ among frequencies	31.6	31.600
none	differ among frequencies	differ among frequencies	32.4	48.200
individual on intercept	global	global	103	103.000
none	global	global	109	125.000

Note that the chosen model included individual participant as random effect, suggest that there exist some level of correlation among the test result of one participant. The model also used grouping based on carrier rather than frequency, suggested that frequency is not enough for accurately predicting EFRs (i.e. EFRs may also depend on first and second volwel formants). However, all 8 carriers share the same intercept term, i.e. the baseline probability of having an EFR detected when $SL = 0$ is probably not related with carriers.

Figure.3 showed the model predicted SL threshold and its corresponding bootstrap 0.95 CI for both methods. Noticed that all CIs are above 0. This may suggest that EFR based method was less sensitive than known SL method (i.e. a positive SL was needed to reach 0.5 chance of detecting an EFR). Further, SL threshold dependents on carriers, though I observed a general trend that high frequency group had lower threshold than other two. Differences on threshold between different carriers needed further investigations. Results in this analysis also showed the difference can be large, in terms of point estimations of F-test, largest difference was between carrier s (6.030 .95CI=(4.08,8.00)) and carrier /a/F1 (33.35, .95CI=(26.12,41.46)) for amplitude method , point estimation difference can be large as 27.32 unit of SL given our surveyed SL has range of (-13.8,54.9). Threshold for detecting an EFR based on phase (Rayleigh-test) was generally similar with those based on amplitude (F-test) when both choose 0.05 as cut off (Figure.3). Numeric point estimation

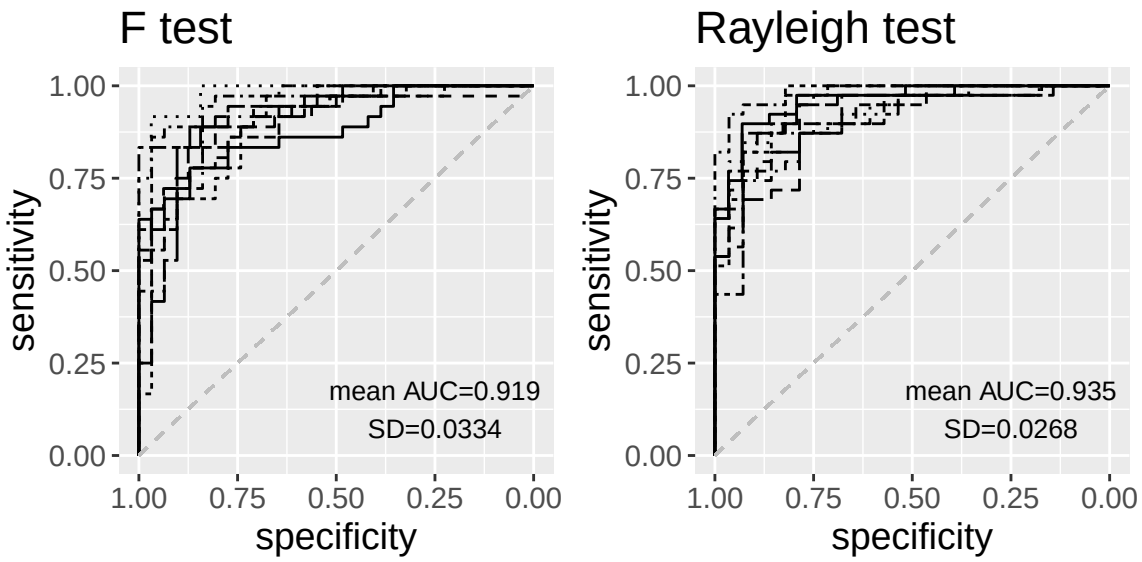

Figure 3: 10-fold corss validation obtained ROC for the mixed effect model predict EFR using SL and carrier with least AIC

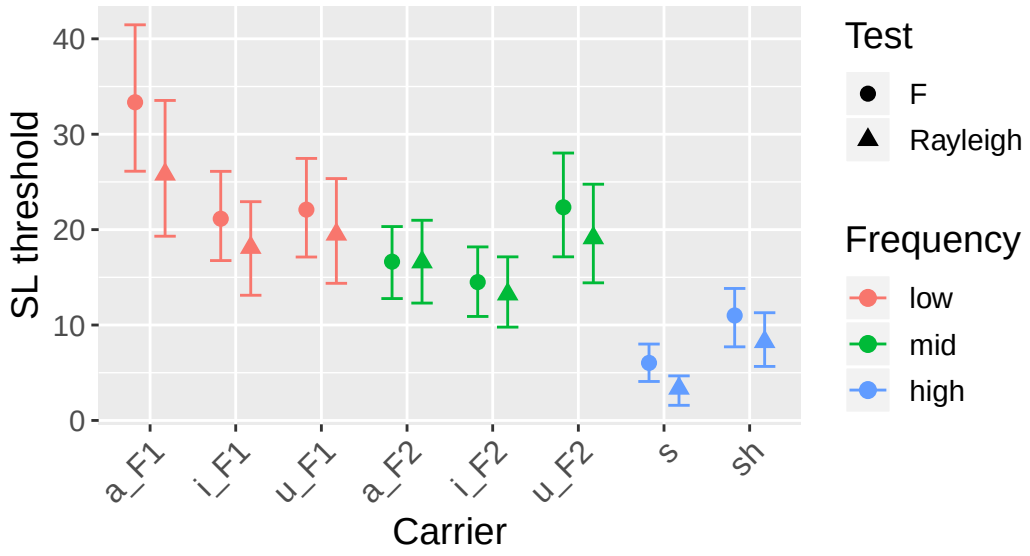

Figure 4: SL thresholds that reach 0.5 chance of detecting EFR according to least AIC model, points: point estimation of the threshold based on 10-fold cross validation, errorbars: 0.95 bootstrap CI

and 0.95 bootstrap CI can be found in appendix Table.5.

The full estimated response curve (i.e. probability of detecting a EFR as a function of SL) based on average prediction of 10-fold cross validation was given in Figure.5.

Discussion

Cut off p-values for Detection

By using the ROC curve, we may find some "better" cut off values than conventional 0.05. For example, cut off value that maximized sum of sensitivity and specificity was 0.0898 for F test and 0.0795 for Rayleigh test. However in reality we may not have the ability to obtain the better cut-off values from EFR data itself. If we want to control the sample bias i.e. more likely to detect an EFR when using carrier `s`, it is beneficial to pre-train the cut-off using benchmark method like SL before conducting further research based on EFR results.

Performance of EFR as an Indicator of Audibility, some Implication for Further Research

Overall, both amplitude (F-test) and phase (Rayleigh-test) based EFR in predicting audibility based on the overall AUC values (0.837 and 0.830). However, when using 0.05 as cut off, both method tend to have relative low sensitivity (0.634 and 0.679). This is possible the cause of a relative poor overall consistent rate (0.688 and 0.717) with traditional SL based method. EFR may not be a good choice when high sensitivity of detection is needed. But relative high specificity may also because the imbalanced experimental design I will cover later.

Performance of both test depended on carriers and their frequency, general trend was carriers that had higher frequency had a lower threshold for a detection thus higher sensitivity. This unbiased nature should be treated with carefulness when comparing the result of EFR from different carriers, as we saw, the difference between thresholds can be large. For same individual, detection threshold for different carriers had correlation. This correlation needs to be accounted when conducting research based on EFR.

Possible Improvement in Analysis and Experimental Design

There were several parts can be improved but relies on experimental design. First of all, I calculated both specificity and sensitivity. On this data set, specificity is relatively high (0.986, 0.974 for F and Rayleigh) when using 0.05 as cut off for EFR detection. This may indicate the method was specific but also can because of the imbalanced design focused on addressing the SL threshold of detecting. There were 560 audible speech based on SL, which is 83.3% of the full sample size. Thus even if the test method gave results that all speeches were audible, specificity is still as high as 0.833. The specificity maybe over estimated. To address this question, researchers should conduct more experiment at the region that SL is negative and test whether an EFR can be detected.

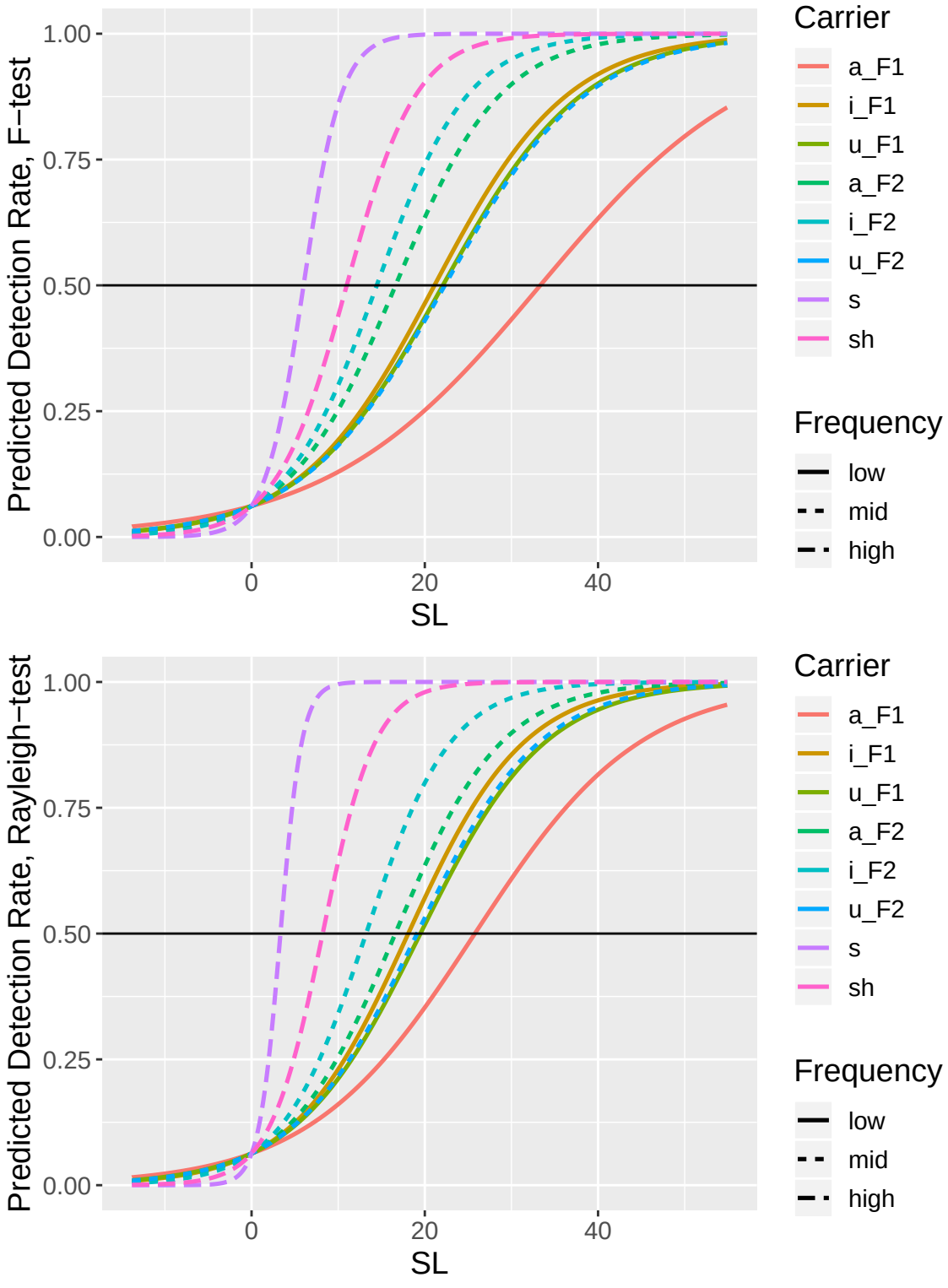

Figure 5: Predicted probability of detecting a EFR as a function of SL, calculate as average of 10 cross validations on logit scale

I used bootstrap for confidence interval and cross validation to make and evaluate the performance of point estimations. However there was no theoretical guarantee in finite sample case for both of the methods. For bootstrap, repeat sampling from a biased sample will still be biased and the empirical sample distribution estimated by bootstrap will be off. The best way of solving this problem was to repeat more trails and assess the experimental design for possible bias on equipment, population tested, etc.

One of the key assumption in the third part of the analysis was different individual's results were independent. This assumption could be violated if all the experiments were done using the same equipment for a given carrier/frequency. This assumption should be assessed based on the experimental design and repeat from other labs.

End of the Main Report

Open Source Statement

All the souce code, including any analysis and souce code generating plots, tables as well as this report, can be found in: https://github.com/YunyiShen/UW-Course-Projects/tree/master/STAT849_Final

As a young scientist, I believe that being open is the best way being honest in analyzing any data set. This does not only include open the data set, the software used, but also the source of reports.

Table 1: Contingency table between consistency (with SL prediction) and carrier, Rayleigh test

	FALSE	TRUE
a_F1	36	48
a_F2	29	55
i_F1	26	58
i_F2	21	63
s	4	80
sh	14	70
u_F1	29	55
u_F2	31	53

Table 2: Contingency table between consistency (with SL prediction) and carrier, F test

	FALSE	TRUE
a_F1	44	40
a_F2	28	56
i_F1	27	57
i_F2	21	63
s	8	76
sh	16	68
u_F1	33	51
u_F2	33	51

Appendix

Contingency tables

There are 4 contifency tables used to test the independency between EFR prediction ability of SL and carrier/frequency. From Table.1-Table.4

Thresholds

Numerical result for threshold are given in Table.5.

Table 3: Contingency table between consistency (with SL prediction) and frequency, Rayleigh test

	FALSE	TRUE
high	18	150
low	91	161
mid	81	171

Table 4: Contingency table between consistency (with SL prediction) and frequency, F test

	FALSE	TRUE
high	24	144
low	104	148
mid	82	170

Table 5: Numerical Estimation of SL thresholds for detecting EFR and their 0.95 bootstrap CIs

carrier	Test	threshold	0.95CI lower	0.95CI upper	Frequency
a_F1	F	33.3494	26.1234	41.4678	low
i_F1	F	21.1418	16.7570	26.1093	low
u_F1	F	22.0963	17.1294	27.4641	low
a_F2	F	16.6418	12.7740	20.3204	mid
i_F2	F	14.4986	10.8986	18.1896	mid
u_F2	F	22.3430	17.1494	28.0283	mid
s	F	6.0303	4.0823	8.0062	high
sh	F	11.0030	7.7158	13.8335	high
a_F1	Rayleigh	25.7979	19.3099	33.5398	low
i_F1	Rayleigh	18.1167	13.1195	22.9303	low
u_F1	Rayleigh	19.5046	14.3725	25.3544	low
a_F2	Rayleigh	16.5957	12.3025	20.9801	mid
i_F2	Rayleigh	13.2327	9.7794	17.1451	mid
u_F2	Rayleigh	19.1218	14.4221	24.7604	mid
s	Rayleigh	3.3578	1.5853	4.6710	high
sh	Rayleigh	8.2408	5.6587	11.2899	high

Cross validation for mix effect model

Cross validation function was not yet implemented in package `lme4` but it is relatively easy to implement with the help of `caret`

```
require(caret)
CV_glmer = function(obj,fold){
   response = as.character( obj@call$formula[2]) # get the name of response
   data_ = obj@frame # get the data
   family_ = as.character( obj@call$family) # get family

   folds = createFolds(y=data_[,response]
                       ,k=fold) # create a fold

   formular = obj@call$formula # get formular

   retrain = lapply(1:fold,function(i,folds,
                                    formular,
                                    data_,
                                    response,
                                    family_){
                            #start ith cross validation

                            # train with training set:
                            retrained_model = glmer(formular,
                                  data = data_[-folds[[i]],],
                                  family = family_)


                            # predict the testing:
                            pred_retrain = predict(retrained_model,
                            newdata = data_[folds[[i]],],
                            type = "response")
                            roc_data =
                               data.frame(resp =
                                             data_[folds[[i]],response],
                                          pred = pred_retrain)
```

```
                                # return the retrained model and prediction:
                                return(list(
                                   model = retrained_model,
                                   roc_data=roc_data))
                },folds,formular,data_,response,family_) # end the lapply

    ## do ROC
    roc_res = lapply(retrain,function(w){
                                roc(w$roc_data,
                                    response = "resp",
                                    predictor = "pred")
       }) # get ROC using the test sample

   return(list(ROC=roc_res,models = retrain)) # return result

}
```

References

Angelo Canty and Brian Ripley (2019). boot: Bootstrap R (S-Plus) Functions. R package version 1.3-23.

Davison, A. C. & Hinkley, D. V. (1997) Bootstrap Methods and Their Applications. Cambridge University Press, Cambridge. ISBN 0-521-57391-2

DeLong, Elizabeth R., David M. DeLong, and Daniel L. Clarke-Pearson. "Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach." Biometrics (1988): 837-845.

Douglas Bates, Martin Maechler, Ben Bolker, Steve Walker (2015). Fitting Linear Mixed-Effects Models Using `lme4`. Journal of Statistical Software, 67(1), 1-48. doi:10.18637/jss.v067.i01.

Max Kuhn. Contributions from Jed Wing, Steve Weston, Andre Williams, Chris Keefer, Allan Engelhardt, Tony Cooper, Zachary Mayer, Brenton Kenkel, the R Core Team, Michael Benesty, Reynald Lescarbeau, Andrew Ziem, Luca Scrucca, Yuan Tang, Can Candan and Tyler Hunt. (2019). `caret`: Classification and Regression Training. R package version 6.0-84. https://CRAN.R-project.org/package=caret