

다변량분석 2번째 과제

# Assignment 2

## Logistic Regression

### ‘목소리 성별 맞추기’

학과  
학번  
이름

산업경영공학부  
2016170863  
추창욱

**[Q1] 본인이 스스로 Logistic Regression 모델을 적용하는 데 적합한 데이터셋을 선정하고 선정 이유를 설명하시오.**

답)



데이터를 찾는 과정에서 가장 먼저 찾았던 데이터는 telecommunication 관련 회사의 고객군을 예측하는 데이터였습니다. 다중 로지스틱 회귀 모델이었는데 정확도가 너무나도 낮아 40%보다 못 미치는 결과값이 나왔습니다. 정확도에 너무 집착하면 안된다는 것을 알지만 너무 낮은 값이 나온 거 같다는 불안감에 쓰지 못하였습니다. 하지만, 그 외에도 유방암 예측, 광고 클릭유무 등과 같이 다양한 데이터를 이용하였으나 만족스러운 결과물을 내지 못했습니다. 여기서 말하는 만족스러운 결과값이 아니었다는 이 데이터셋이 로지스틱 회귀 분석을 하기에 적합하지 않은 데이터 셋이라는 판단을 내렸다는 것입니다.

그러던 와중 수업시간에 교수님께서 얼굴사진을 보고 여성인지 남성인지를 맞춰보던 예제가 생각이 났고 이와 유사하게 목소리의 여러 요소를 통해 목소리의 성별을 예측해 볼 수 있는 데이터 셋을 찾을 수 있었습니다. 데이터를 열어본 순간 여러 소리의 파동과 최고 발산 에너지 등등의 요소를 통해 여성인지 남성인지를 맞추는 데이터라는 것을 알 수 있었고, 종속변수가 성별이라는 점에서 binary logistic regression이 되겠다는 생각이 들었습니다. 분석을 시작하기 전부터 이 변수는 영향을 줄 것 같은데? 이 변수는 당연히 연관을 안 줄 거 같은데? 와 같은 가설들을 스스로 내리게 되었고, 목소리는 사람을 성별을 구분하는데 있어 상당히 뚜렷한 요소 중 하나인데, 여기 있는 변수들은 전부 목소리의 속성과 관련된 변수들이고 당연히 높은 예측률이 나올 것 같지만 반대로 낮은 예측률이 나오면 그것도 나름대로 정말 흥미로울 것 같다 라는 생각이 가장 많이 들었던 것 같습니다.

데이터 다운 링크: <https://www.kaggle.com/primaryobjects/voicegender>

**[Q2] 해당 데이터셋의 종속 변수와 설명 변수는 어떤 것 들이 있는가? 분석 전에 아래 두 가지 질문에 대해서 스스로 생각해 보고 답변을 하시오.**

우선, 데이터 분석을 진행하기 전에 어떤 변수들이 존재하는지에 대해 알아볼 필요가 있다고 판단하여 다음과 같이 정리를 해보았습니다.

변수명	설명
meanfreq	평균 주파수(kHz 단위)
sd	주파수의 표준 편차
median	중위수 주파수(kHz 단위)
Q25	첫 번째 분위수(kHz 단위)
Q75	세 번째 분위수(kHz 단위)
IQR	양자 간 범위(kHz)
skew	왜도
kurt	첨도
sp.ent	스펙트럼 엔트로피
sfm	스펙트럼 평탄도
mode	모드 주파수
centroid	주파수 중심
meanfun	음향 신호 전반에서 측정된 기본 주파수의 평균
minfun	음향 신호 전반에 걸쳐 측정된 최소 기본 주파수
maxfun	음향 신호 전반에 걸쳐 측정된 최대 기본 주파수
meandom	음향 신호 전반에서 측정된 우성 주파수의 평균
mindom	음향 신호 전반에 걸쳐 측정된 최소 우성 주파수
maxdom	음향 신호 전반에서 측정된 최대 우성 주파수
dfrange	음향 신호 전반에 걸쳐 측정된 지배적인 주파수 범위
modindx	변조 인덱스로서 기본 주파수를 주파수 범위로 나눈 인접 주파수 측정 간의 누적 절대 차이로 계산됨
label	남성 또는 여성

라벨이라는 변수는 여성 또는 남성임을 나타내는 변수이므로 종속 변수가 될 것이며, 이 종속 변수 즉, 성별을 가리는데 사용되는 설명 변수는 하얀색 row들일 것입니다.

**1. 이 데이터에서 제공된 설명 변수들 중에서 높은 상관 관계가 있을 것으로 예상되는 변수들은 어떤 것들이 있는가? 왜 그렇게 생각하는가?**

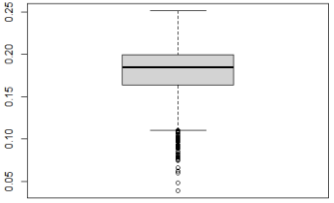
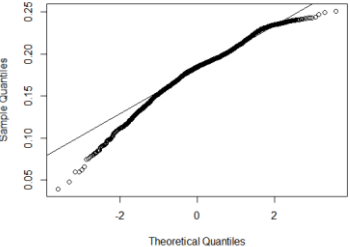
답) 변수들 간에 상관관계가 높을 것이라고 예상되는 변수는 meanfreq-meanfun이었습니다. 그 이유는 두 변수 모드 주파수의 평균이라는 정보를 내포하고 있는데 음향 신호 전반에서 측정된 기본 주파수와 전체 주파수가 크게 다를 것 같지 않다는 생각을 하였기 때문입니다. 그 다음은 왜도와 첨도였습니다. 왜도와 첨도는 주파수의 파동의 형태를 나타내는 변수인데, 보통 주파수는 목소리에 한하여 높은 주파수를 가지면 왜도와 첨도가 동시에 줄어들고, 낮은 주파수에 대해서는 왜도 첨도가 동시에 증가하기 때문에 매우 큰 상관관계가 있을 것이라고 생각 하였습니다. 마지막으로 sp.ent와 sfm이 높은 상관관계를 가질 것이라고 예상하였습니다. 왜냐하면 간단하게 말해서 둘 다 스펙트럼에 관한 변수이기 때문입니다.

## 2. 제공된 설명 변수들 중에서 종속 변수를 예측하는 데 필요하지 않을 것으로 예상되는 변수들은 어떤 것 들이 있는가? 왜 그렇게 생각하는가?

답) 가장 먼저 필요가 없을 거라고 생각한 데이터는 **sd(주파수 표준편차)**, **sp.ent(스펙트럼 엔트로피)**, **sfm(스펙트럼 평탄도)**, **dfrange** 입니다. 우선 sd와 같은 경우는 1번 에서 언급한 바와 같이 주파수는 소리의 높낮이와 소리 크기 음색 등등을 나타내는 소리의 요소 인데 이 표준편차가 의미하는 바는 소리의 폭이 얼마나 넓냐 이지 여성과 남성을 판가름 할 수 있는 변수는 아닐 거라는 생각이 들었습니다. 또한 소리는 에너지의 이동이지만, 제가 조사한 바에 따르면 스펙트럼의 활용분야로 대표적인 것이 빛, 전기 등이라는 점에서 소리와는 크게 상관이 없을 거라는 예상을 해보았습니다. Dfrange는 어떤 것을 의미하려는 건지는 사실 잘 모르겠으나, 누적 절대 차이라는 점에서 상관성이 정말 없을 것 같다는 생각을 하였습니다. 그 이유는 절대값을 붙인다는 것은 음수이든 양수이든 별 상관이 없고 그 값의 크기만을 본다는 것인데, 과연 이러한 데이터가 여성인지 남성인지를 판가름 하는데 있어 유의미한 영향을 끼칠 수 있는 것이 있을까? 라는 의문점이 들었기 때문입니다.

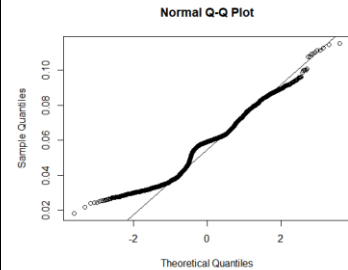
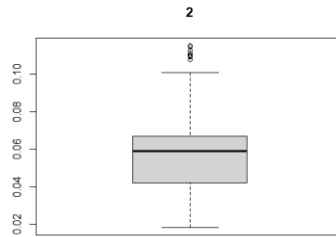
**[Q3] 모든 연속형 숫자 형태를 갖는(명목형 변수제외) 개별 입력 변수들에 대하여 각각 다음과 같은 단변량 통계량을 계산하고 Box plot을 도시하시오: Mean, Standard deviation, Skewness, Kurtosis. 전체 변수 중에서 정규분포를 따른다고 할 수 있는 변수들은 몇개인가? 정규분포를 따른다고 가정한 근거는 무엇인가?**

답) 3번 질문에 대해 답을 하기 위해 반복문을 통해 각 변수(칼럼) 들에 대한 평균, 표준편차, 왜도, 첨도를 구해보았습니다. 이때에 왜도는 0에 가까울수록, 첨도는 3에 가까울수록 정규분포에 가깝다고 알려져 있으며, 평균과 표준편차를 통해 어떤 분포의 형태를 띄게 될지 알아봄과 동시에 qqplot을 통해 정규성을 확인해보는 과정을 거쳐보았습니다.

단변량 통계량 계산 및 분석		
평균, 표준편차, 첨도, 왜도	박스플롯	Q-Q plot
<b>1 번 칼럼</b> Mean: 0.180907 Standard deviation: 0.029918 Skewness: -0.617203 Kurtosis: 3.801997		

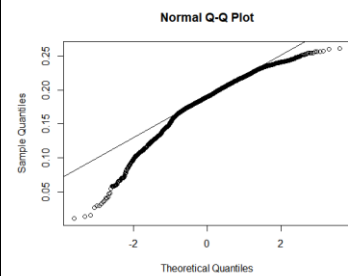
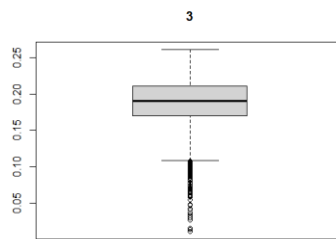
첫번째 변수는 평균 주파수를 의미하며, 왜도와 척도는 정규분포의 형태에서 조금 벗어난듯한 경향이 있으며 박스 플롯을 보면 이상치가 꽤나 존재한다는 것을 확인 할 수 있습니다. Q-Q plot 또한 그다지 안정적인 정규분포를 띠는다고는 할 수 없을 것 같습니다.

2 번 칼럼  
Mean: 0.057126  
Standard deviation: 0.016652  
Skewness: 0.136851  
Kurtosis: 2.477141



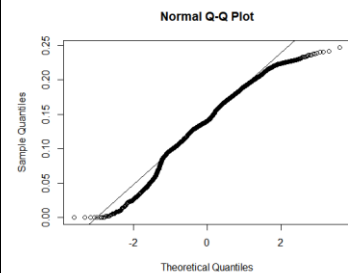
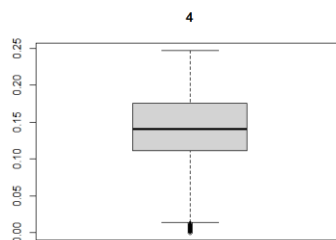
2번째는 주파수의 표준편차인데 아무래도 위와 비슷한 이유에 의해서 정규성을 만족하지 못한다고 할 수 있을 것 같습니다.

3 번 칼럼  
Mean: 0.185621  
Standard deviation: 0.036360  
Skewness: -1.012305  
Kurtosis: 4.625037



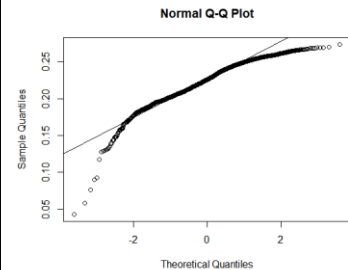
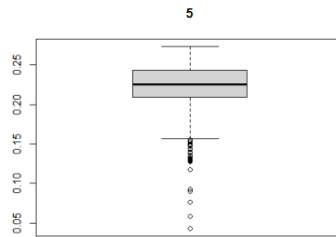
세번째는 중위수 주파수 입니다. 주파수에는 영역이라는 개념이 있는데 주파수를 독립변수로 하는 영역을 말하며 시간 영역에 대응됩니다. 그중 중위수 주파수를 나타내는 변수인데 이 또한 정규성을 만족하는 데이터라기에는 왜도와 첨도가 많이 벗어나 있고 이상치가 존재하며 Q-Q plot 또한 선에 근접해 있지 않다는 점에서 정규성을 만족하지 않는다고 볼 수 있을 것 같습니다.

4 번 칼럼  
Mean: 0.140456  
Standard deviation: 0.048680  
Skewness: -0.490644  
Kurtosis: 3.016411



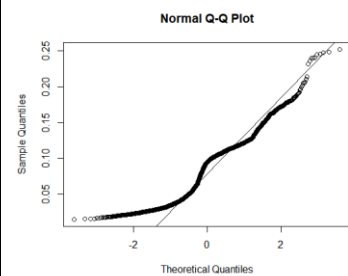
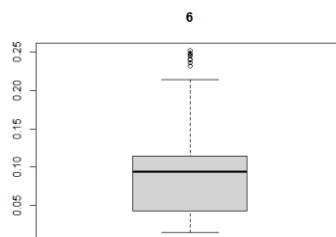
4번째는 첫번째 분위수를 의미하는데 왜도와 척도가 나름 정규 분포에 근사하며 박스플롯과 Q-Q plot 을 봤을 때 살짝 엇그나긴 하지만 충분히 정규 분포를 따른다고 할 수 있을 것 같다고 생각하였습니다.

5 번 칼럼  
Mean: 0.224765  
Standard deviation: 0.023639  
Skewness: -0.899884  
Kurtosis: 5.975213



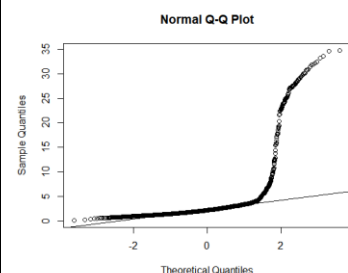
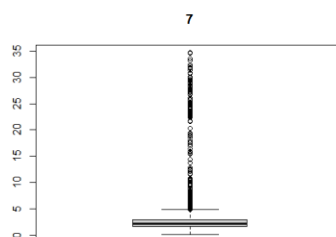
5번째는 세번째 분위수를 의미하는데 왜도와 척도의 값을 보았을 때, 정규분포를 따른다고 하기에는 너무 값이 벗어나 있다는 것을 알 수 있었습니다.

6 번 칼럼  
Mean: 0.084309  
Standard deviation: 0.042783  
Skewness: 0.295292  
Kurtosis: 2.550653

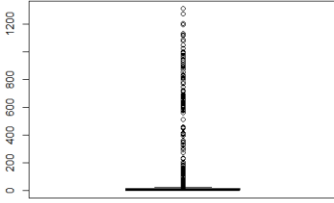
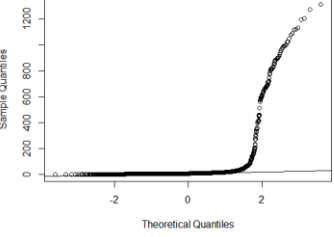
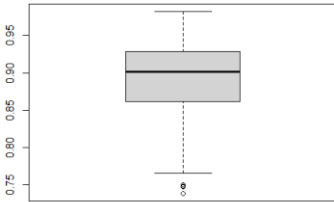
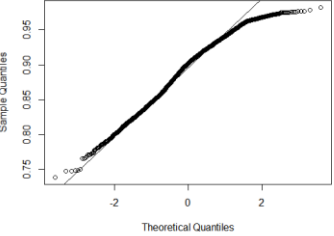
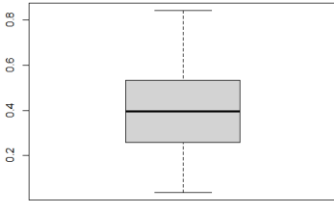
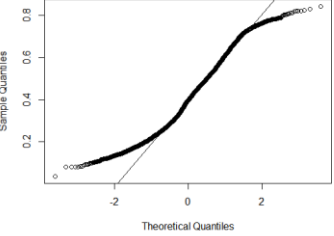


6번째는 양자 간 범위를 의미하는 변수입니다. 박스 플롯과 Q-Q plot을 보았을 때는 정규분포를 따른다고 하기에는 무리가 있을 것 같아 보이지만, 왜도와 첨도가 제가 정한 기준(0.5 사이)에 있으므로 정규분포와 비슷한 형태를 띄고 있을 것이라는 생각을 하게 되었습니다.

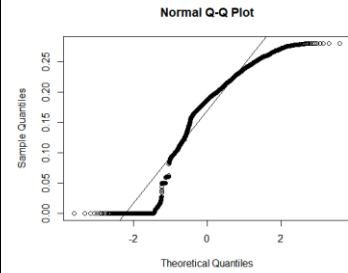
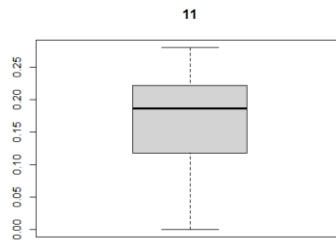
7 번 칼럼  
Mean: 3.140168  
Standard deviation: 4.240529  
Skewness: 4.930978  
Kurtosis: 28.321540



역시나 예상했던 것과 마찬가지로 이 변수는 skewness라는 변수인데 이미 데이터의 왜도를 구해놓은 데이터 변수라는 생각이 들었습니다. 정규분포를 따르지 않는다는 것이 명확했습니다.

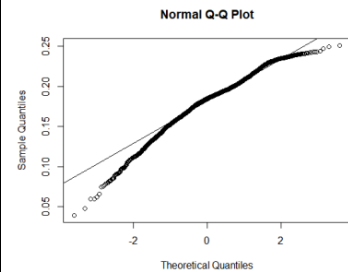
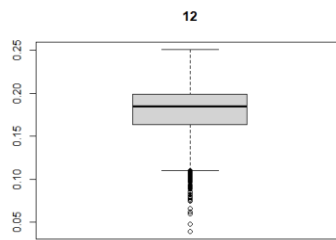
<p>8 번 칼럼</p> <p>Mean: 36.568461 Standard deviation: 134.928661 Skewness: 5.869805 Kurtosis: 38.873550</p>	<p>8</p> 	<p>Normal Q-Q Plot</p> 
<p>위와 같은 형태의 변수로 kurtosis변수입니다. 이미 데이터의 첨도를 구해놓은 데이터 변수인데 첨도의 첨도를 구한것이라 당연히 정규분포를 따르지 않겠구나 라고 생각하였습니다. 정규분포를 따르지 않는다는 것이 명확했습니다.</p>		
<p>9 번 칼럼</p> <p>Mean: 0.895127 Standard deviation: 0.044980 Skewness: -0.430730 Kurtosis: 2.574851</p>	<p>9</p> 	<p>Normal Q-Q Plot</p> 
<p>9번 변수는 스펙트럼 엔트로피를 보여주는 변수인데 왜도와 첨도 0과 3에서 0.5 이내의 값을 가지며 박스 플롯과 Q-Q plot이 정규성을 잘 보여주고 있다는 판단을 내릴 수 있었습니다.</p>		
<p>10 번 칼럼</p> <p>Mean: 0.408216 Standard deviation: 0.177521 Skewness: 0.339797 Kurtosis: 2.163492</p>	<p>10</p> 	<p>Normal Q-Q Plot</p> 
<p>10번 변수는 스펙트럼의 평탄도 인데 박스 플롯과 Q-Q plot은 상당히 정규 분포에 흡사한 형태를 띤다고 보여지나 첨도가 0.5 범위 밖인 것을 보아 일반 정규 분포보다는 높이가 좀 낮은 형태의 분포를 띤 것이라는 것을 알 수 있었습니다.</p>		

11 번 칼럼  
Mean: 0.165282  
Standard deviation: 0.077203  
Skewness: -0.836840  
Kurtosis: 2.742603



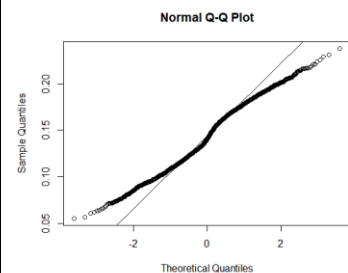
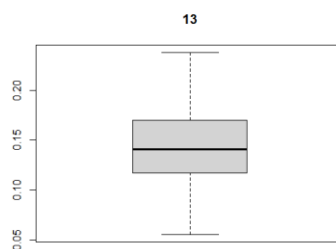
모드 주파수는 오른쪽으로 치우친 형태의 분포를 가지고 있으며 Q-Q plot을 보면 도중에 굽기 구간이 보이는데 데이터들의 분포가 특정 구간에서 상당히 적다는 것을 알 수 있다는 것을 확인 할 수 있어 정규분포를 따른다 기에는 부족하다고 판단하였습니다.

12 번 칼럼  
Mean: 0.180907  
Standard deviation: 0.029918  
Skewness: -0.617203  
Kurtosis: 3.801997



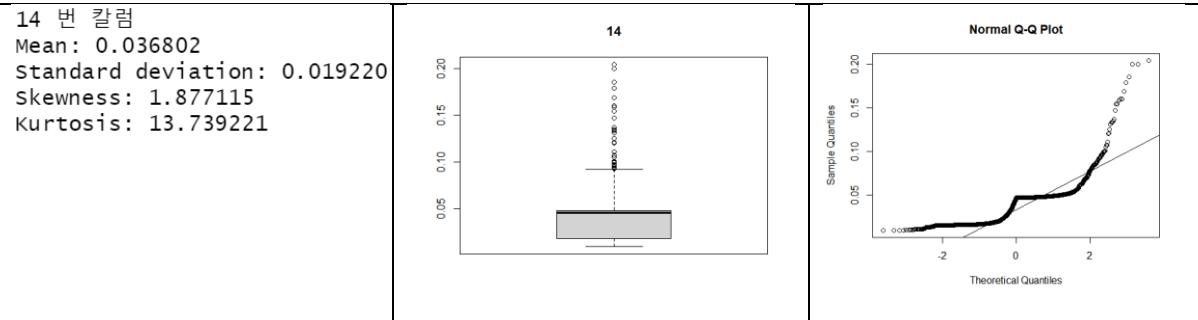
주파수 중심을 나타내는 변수인데 오른쪽으로 좀 치우친 경향이 있으며 높이 또한 높은 형태의 분포를 띠는 것을 알 수 있었습니다. 결론적으로는 정규분포를 만족한다 기에는 다소 무리가 있지 않을까 라는 생각을 하였습니다.

13 번 칼럼  
Mean: 0.142807  
Standard deviation: 0.032304  
Skewness: 0.039122  
Kurtosis: 2.139504

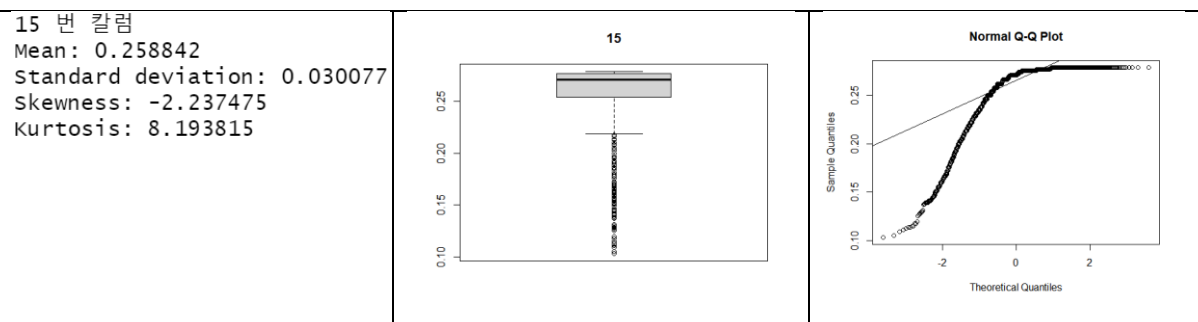


기본 주파수의 평균으로서 첨도와 Q-Q plot을 제외하고는 정규분포의 형태와 유사하는 것을 확인 할 수 있었습니다. 하지만 상대적으로 정규 분포와 유사한 형태를 띄고 있을 것이라고 예상하였는데 그 이유는 Q-Q plot에서 선에서 멀어지는 데이터들의 거리가 비교적 일정하다는 점 때문이었습니다.

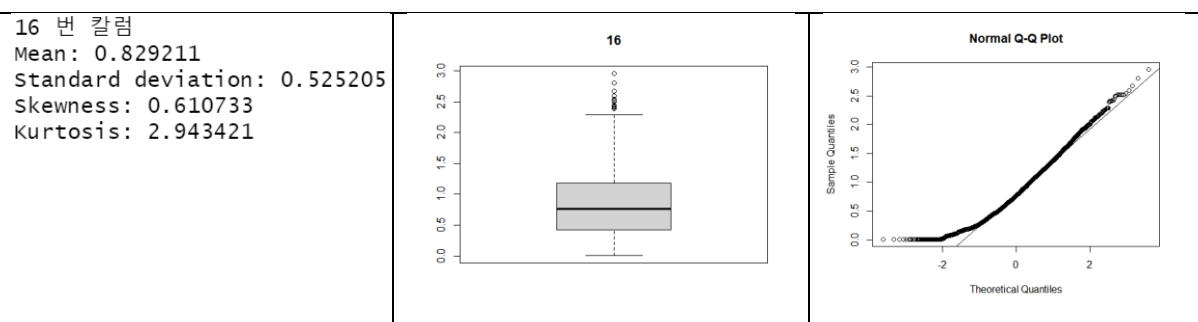




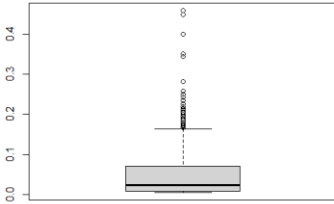
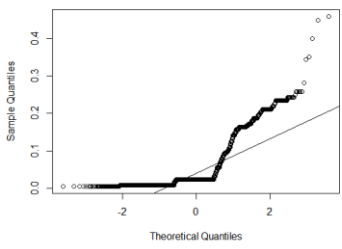
왜도와 첨도 뿐만 아니라 데이터의 분포가 전반적으로 상당히 제각각이고 일정하게 분포되어있지 않다는 점에서 정규분포를 따르지 않는다는 결론을 내렸습니다.



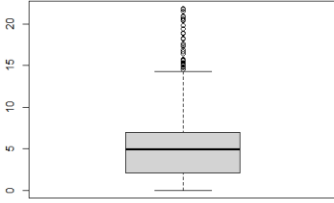
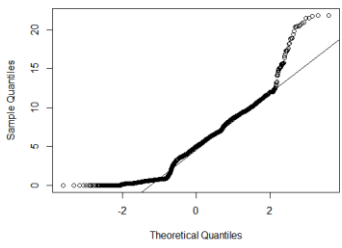
이 변수는 maxfun에 해당하는 변수인데 역시나 위의 minfun과 비슷하게 정규 분포와는 다른 형태라는 것을 알 수 있었습니다. 왜 그런가에 대해 생각을 해보았을 때, 쉽게 생각하여 개인이 낼 수 있는 목소리 음역대의 최대치와 최소치를 기록한 데이터인데 이는 일반적인 데이터가 아닌 데이터의 성질자체가 최소와 최대를 기록한 데이터이므로 데이터의 분포가 상당히 차이날 수 밖에 없다는 생각이 들었습니다.



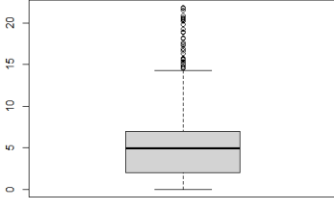
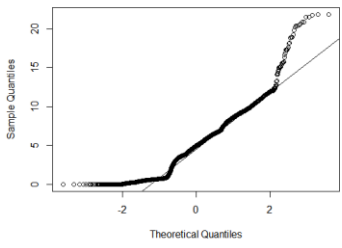
전반에서 측정된 우성 주파수의 평균을 나타내는 이 데이터는 다른 조건들을 정규 분포와 근사하라는 결론을 내릴 수 있는 조건을 갖추고 있지만 왜도가 너무 큰 값을 가집니다. 즉 왼쪽으로 치우쳐 있는 형태를 띠는 점에서 정규분포와 비슷한 형태가 아니라는 결론을 내릴 수 있습니다.

17 번 칼럼 Mean: 0.052647 Standard deviation: 0.063299 Skewness: 1.660327 Kurtosis: 5.182242		
---	---	---

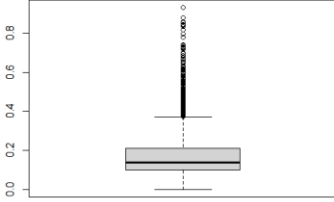
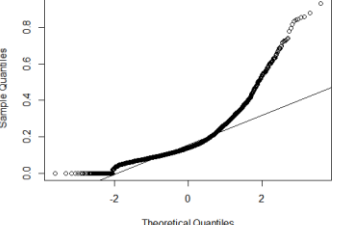
최소 우성 주파수를 나타내는 이 그래프와 같은 경우 왜도 첨도 박스 플롯, Q-Qplot 모두 정규 분포를 따르지 않는다는 것을 보여주고 있으므로 정규분포를 따르지 않는다는 결론을 내릴 수 있었습니다

18 번 칼럼 Mean: 5.047277 Standard deviation: 3.521157 Skewness: 0.725845 Kurtosis: 4.310770		
---	--	--

최대 우세 주파수 또한 위와 같은 이유에서 정규 분포를 따르지 않는다는 결론을 내릴 수 있습니다.

19 번 칼럼 Mean: 4.994630 Standard deviation: 3.520039 Skewness: 0.727916 Kurtosis: 4.314040		
---	---	---

지배적인 주파수의 범위를 나타내는 이 변수의 경우도 첨도와 왜도가 너무 컸다가 있으며 이상치도 많은 탓에 정규 분포를 따른다고 하기에 어려움이 있을 것으로 예상됩니다.

<p>20 번 칼럼</p> <p>Mean: 0.173752</p> <p>Standard deviation: 0.119454</p> <p>Skewness: 2.063357</p> <p>Kurtosis: 8.913695</p>		
<p>이는 변조 인덱스입니다. 한마디로 어떤 특정한 함수나 결과를 위해 가공한 데이터라는 뜻으로 왜도와 첨도 박스플롯 그리고 Q-Q plot에서도 알 수 있다시피 정규분포를 따르지 않는다는 것을 확인할 수 있었습니다.</p>		

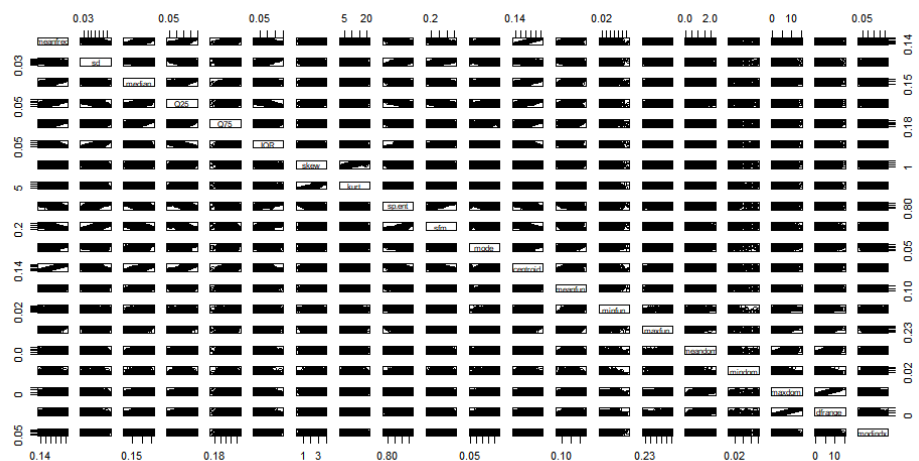
**[Q4] [Q3]의Box plot을 근거로 각 변수들에 대한 이상치(너무 값이 크거나 작은 객체) 조건을 정의하고, 해당하는 객체들을 데이터셋에서 제거 해보시오**

답) 위의 결과를 보고 가장 먼저 알 수 있는 결과는 이상치가 정말 많다는 것입니다. 통계적으로는 상/하위 3 표준편차를 벗어나거나 1.5 IQR을 벗어나면 극단치 또는 이상치로 판단한다고 합니다. 이번 과제에서는 각 변수의 1분위수 0.25와 3분위수 0.75에서 1.5 IQR를 벗어나는 데이터들을 제거하여 보았습니다. 코드는 상한선과 하한선을 정하듯 기준선을 넘어가는 행을 삭제하도록 설정하여 코드를 실행해주었습니다.

**[Q5] 가능한 모든 두쌍의 입력 변수 조합에 대한 산점도(scatter plot) 및 Correlation plot (hint: "corrplot" 패키지의corrplot( ) 함수사용) 상관관계를 계산해보시오.**

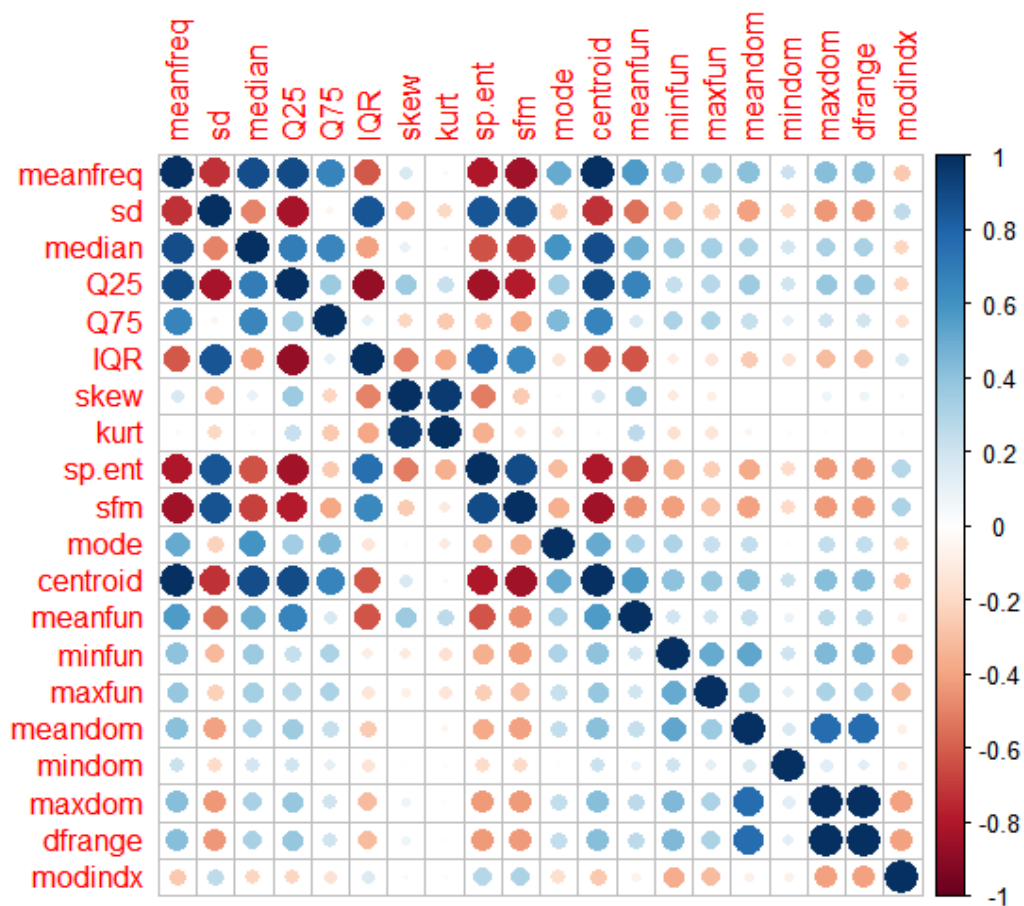
**1. 어떤 두 조합의 변수들이 서로 강한 상관 관계가 있다고 할 수 있는가?**

답) 가장 먼저 실행한 작업은 1~20 변수들의 scatter plot을 확인해 보는 것이었습니다. 하지만 20개가 정도의 변수가 존재하고 가시성이 너무 떨어져 어떤 변수들끼리 강한 상관 관계가 있는 것인지 파악을 할 수가 없었습니다. 결과물은 아래와 같았습니다.



그래서 진행하게 된 방법은 더 보기 좋은 `corrplot()`을 이용하자는 것이었습니다. 결과물은 아래와 같았습니다.

우선 주목해서 봐야할 색깔은 역시나 진한 파란색원이 존재하는 두 변수였습니다. **Meanfreq-centroid, kurt-skew, dfrange-maxdom** 등의 관계가 대표적으로 아주 높은 양의 상관 관계를 띠는 것을 확인할 수 있었습니다. 또한, **Q25-sd, sfm-meanfreq, IQR-Q25, centroid-sfm**은 아주 강한 음의 상관관계를 가지는 것을 확인 할 수 있었습니다. 그림으로만 분석하기에는 너무 어림 잡아 해석을 하는 것 같아 수치를 통해 다시 확인을 해보았습니다.



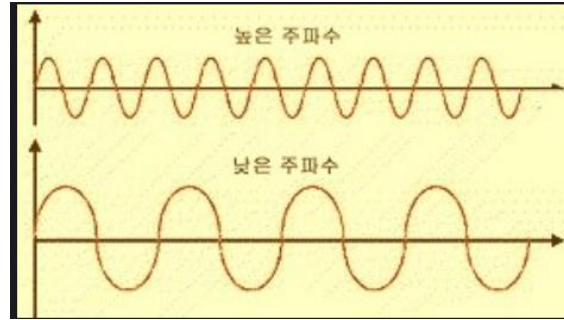
	meanfreq	sd	median	Q25	Q75	IQR	skew	kurt	sp.ent	sfm	mode
meanfreq	1.00000000	-0.72949820	0.88228876	0.8925682	0.66940551	-0.61076600	0.16367334	0.03150399	-0.8037324	-0.8476381	0.50740672
sd	-0.72949820	1.00000000	-0.49368785	-0.8260822	-0.05596388	0.85306059	-0.32805968	-0.20472928	0.8588630	0.8616971	-0.22557878
median	0.88228876	-0.49368785	1.00000000	0.6932636	0.65224877	-0.40681865	0.09395046	-0.03214222	-0.6345165	-0.6814455	0.59944281
Q25	0.89256824	-0.82608219	0.69326360	1.00000000	0.36678555	-0.87995474	0.36309919	0.22987830	-0.8303762	-0.7873118	0.34470774
Q75	0.66940551	-0.05596388	0.65224877	0.36678555	1.00000000	0.11919396	-0.21448071	-0.26658199	-0.2654810	-0.3838730	0.44934740
IQR	-0.61076600	0.85306059	-0.40681865	-0.87995474	0.11919396	1.00000000	-0.49704247	-0.38146745	0.7506542	0.6442371	-0.13843242
skew	0.16367334	-0.32805969	0.09395046	0.3630992	-0.21448071	-0.49704245	1.00000000	0.95286849	-0.5197391	-0.2585112	0.02317763
kurt	0.03150399	-0.20472928	-0.03214222	0.2298783	-0.26658199	-0.38146745	0.95286849	1.00000000	-0.3585965	-0.1173772	-0.11316487
sp.ent	-0.80373240	0.85886304	-0.63451648	-0.8303762	-0.26548099	0.75065417	-0.51973908	-0.35859654	1.00000000	0.8957660	-0.31280988
sfm	-0.84763812	0.86169711	-0.68144555	-0.7873118	-0.38387300	0.64423710	-0.25851153	-0.11737724	0.8957660	1.00000000	-0.35012068
mode	0.50740672	-0.22557878	0.59944281	0.3447077	0.44934740	-0.13843242	-0.02317763	-0.11316487	-0.3128099	-0.3501207	1.00000000
centroid	1.00000000	-0.72949820	0.88228876	0.8925682	0.66940551	-0.61076600	0.16367334	0.03150399	-0.8037324	-0.8476381	0.50740672
meanfun	0.56410384	-0.54343693	0.48674249	0.36958714	0.33911588	0.16213848	0.36996630	0.26529897	-0.6248801	-0.4592394	0.31459932
minfun	0.40756409	-0.32065515	0.36958714	0.2362679	0.16213848	0.16213848	-0.09167798	-0.11344716	-0.16700806	-0.3542974	0.30899124
maxfun	0.38235143	-0.23268523	0.33911588	0.2734982	0.31682012	0.31682012	-0.13010852	-0.087133825	-0.13347649	-0.2462151	0.22754798
meandom	0.41893926	-0.40847810	0.31614540	0.3549700	0.23333425	-0.25969129	-0.001191163	-0.05247688	-0.3703326	-0.4052868	0.24343843
mindom	0.21034683	-0.18360943	0.18676821	0.1915392	0.11242523	-0.14701115	-0.005345192	-0.02837041	-0.1862138	-0.1938839	0.01835621
maxdom	0.42619409	-0.43600954	0.32107912	0.3883283	0.20082921	0.31189155	0.073964578	0.01069647	-0.4234393	-0.4248769	0.24638493
dfrange	0.42561940	-0.43553778	0.32054208	0.3878051	0.20050946	-0.31149645	0.074018773	0.01080582	-0.4229521	-0.4243622	0.24643245
modindx	-0.26998128	0.25822909	-0.21034419	-0.2167206	-0.15131613	0.15402650	-0.034841929	0.02875090	0.2808430	0.3130205	-0.17580236

	centroid	meanfun	minfun	maxfun	meandom	mindom	maxdom	dfrange	modindx
meanfreq	1.00000000	0.56410384	0.40756409	0.38235143	0.41893926	0.21034683	0.42619409	0.42561940	-0.26998128
sd	-0.72949820	-0.54343693	-0.32065515	-0.23268523	-0.40847809	-0.18360942	-0.43600954	-0.43553778	0.25822909
median	0.88228876	0.48674249	0.36958714	0.33911588	0.31614539	0.18676821	0.32107912	0.32054208	-0.21034419
Q25	0.89256824	0.66676705	0.23626792	0.27349823	0.354970018	0.191539218	0.38832825	0.38780506	-0.21672058
Q75	0.66940551	0.16213848	0.31426726	0.31682012	0.233334249	0.112425226	0.20082921	0.20050946	-0.15131613
IQR	-0.61076600	-0.62881338	-0.09167798	-0.13010852	-0.259691292	-0.147011149	-0.31189155	-0.31149645	0.15402650
skew	0.16367334	0.36996631	-0.11344717	-0.08713383	-0.001191163	-0.005345192	0.07396458	0.07401877	-0.03484193
kurt	0.03150399	0.26529897	-0.16700806	-0.13347649	-0.052476885	-0.028370411	0.01069647	0.01080582	0.02875090
sp.ent	-0.80373240	-0.62488015	-0.35429736	-0.24621509	-0.370332578	-0.186213752	-0.42343928	-0.42295207	0.28084295
sfm	-0.84763812	-0.45923940	-0.41210445	-0.29556122	-0.405286766	-0.193883871	-0.42487695	-0.42436220	0.31302052
mode	0.50740672	0.31459932	0.30899124	0.22754798	0.243438431	0.018356206	0.24638493	0.24643245	-0.17580236
centroid	1.00000000	0.56410384	0.40756409	0.38235143	0.418939264	0.210346837	0.42619409	0.42561940	-0.26998128
meanfun	0.56410384	1.00000000	0.19750404	0.19563280	0.231186936	0.081338969	0.26613136	0.26595644	-0.06749664
minfun	0.40756409	0.19750404	1.00000000	0.50117533	0.523093484	0.203539568	0.44226631	0.44172416	-0.36658932
maxfun	0.38235143	0.19563280	0.50117533	1.00000000	0.362747809	0.119511090	0.31308906	0.31279567	-0.31224613
meandom	0.41893926	0.23118694	0.52309348	0.36274781	1.000000000	0.164556091	0.76870706	0.76846071	-0.08519169
mindom	0.21034684	0.08133897	0.20353957	0.11951109	0.164556091	1.000000000	0.12875010	0.12513210	-0.07614580
maxdom	0.42619409	0.26613136	0.44226631	0.31308906	0.768707055	0.128750102	1.000000000	0.99999335	-0.40195536
dfrange	0.42561940	0.26595644	0.44172416	0.31279567	0.768460712	0.125132098	0.99999335	1.00000000	-0.40186296
modindx	-0.26998128	-0.06749664	-0.36658932	-0.31224613	-0.085191692	-0.076145797	-0.40195536	-0.40186296	1.00000000

수치로도 확인을 할 수 있듯 이 변수들 사이의 양의 상관관계는 0.95를 넘는 수치를 가지고 있다는 것을 확인할 수 있었습니다. 음의 상관관계들 또한 -0.85보다 아래의 수치를 가지고 있었습니다. 특히나 양의 상관관계를 가지는 변수들은 1에 육박하는 수치이므로 결코 무시할 수 없는 수치라고 생각하여 이 변수들은 강한 상관관계를 가진다고 생각하였습니다.

## 2.강한 상관 관계가 존재하는 변수 조합들 중에 대표 변수를 하나씩만 선택해서 전체 변수의 개수를 감소시켜보시오 ([Q7]에서사용함)

답) 변수 조합을 통해 대표 변수 하나로 간추리기 위해서는 변수에 대한 이해도를 필요로 합니다. 현재 위에서 거론된 변수들은 **Meanfreq-centroid, kurt-skew, dfrange-maxdom**인데, 우선 meanfreq는 평균 주파수를 의미하고 centroid는 주파수 중심을 의미합니다. 가장 먼저 드는 생각은 주파수의 평균과 주파수 중심이 높은 상관관계를 갖는다는 것은 주파수는 언제나 일정하게 흔들리는가 였습니다. 주파수가 일정성이 있기 때문에 평균과 주파수의 중심이 거의 비슷한 값을 가질 수 있다는 것이 설명된다고 생각하였습니다. 조사 결과, 주파수는 전파가 움직이기는 보이지 않는 하나의 길과 같은 것이라고 합니다.



즉 높은 주파수는 파장의 길이가 짧아지는 것이고 진폭 또한 낮아지는 것인데, 물결 모양으로 비교적 일정하게 움직인다는 것을 확인 할 수 있었습니다. 즉, 주파수의 평균과 중심은 거의 같은 변수일 것이라고 생각하였습니다. 하지만 위에서는 언급을 하지 않았지만 median 변수와 Q25 변수도 주파수의 중심과 비슷한 성질을 가지고 있어 **centroid, median Q25 → meanfreq** 하나의 변수로 줄여도 될 것이라는 생각을 하였습니다.

다음은 skew-kurt 변수입니다. 하지만 이는 저도 잘 알고 있는 왜도와 첨도의 개념인데 주파수의 형태를 보면 위아래로 굴곡이 있는 것을 확인 할 수 있습니다. 하지만 주파수의 성질을 보면 높은 주파수는 왜도와 첨도가 동시에 줄어들고, 낮은 주파수는 왜도와 첨도가 동시에 늘어난다는 것을 확인 할 수 있습니다. 이를 통해 이 변수는 하나만 사용해도 높은 주파수 인지, 낮은 주파수 인지를 확인 할 수 있기에 둘 중 하나인 **skew**로 줄여도 될 것이라는 판단을 하였습니다.

마지막은 dfrange와 maxdom입니다. 이 변수들의 설명을 보면 maxdom은 최대 우세 주파수를 의미하고, dfrange는 지배적인 주파수의 범위를 의미합니다. 상식적으로 생각해보아도 우세적으로 많은 주파수와 지배적으로 많은 주파수의 범위는 거의 똑같은 정보를 내포하고 있다고 봐도 무방하다는 생각을 하였습니다. 결론적으로 이 두 변수를 **maxdom** 하나의 변수로 줄여도 된다는 판단을 할 수 있었습니다.

[Q6] 전체 데이터셋을 70%의 학습 데이터와 30%의 테스트 데이터로 분할한 후 모든 변수를 사용하여 Logistic Regression 모델을 학습해 보시오. 이 때 70:30으로 구분하는 random seed를 저장하시오.

1. 유의수준 0.05에서 유의한 변수의 수는 몇 개인지 확인하고 각 변수들이 본인의 상식 선에서 실제로 유효하다고 할 수 있는지 판단해 보시오.

```
call:
glm(formula = voice_target ~ ., family = binomial, data = voice_trn)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-3.0144	-0.0195	0.0031	0.0943	4.6926

Coefficients: (3 not defined because of singularities)

	Estimate	Std. Error	z value	Pr(> z )	
(Intercept)	-9.7757	6.1687	-1.585	0.113031	
meanfreq	6.6131	3.2398	2.041	0.041233	*
sd	1.1014	1.4297	0.770	0.441094	
median	-1.8396	1.0541	-1.745	0.080934	.
Q25	-5.8389	1.4054	-4.155	3.26e-05	***
Q75	0.3777	1.0718	0.352	0.724548	
IQR	NA	NA	NA	NA	
skew	13.9348	7.6877	1.813	0.069895	.
kurt	-54.9906	37.7585	-1.456	0.145289	
sp.ent	2.4855	1.2120	2.051	0.040287	*
sfm	-1.3321	0.8794	-1.515	0.129830	
mode	-0.1682	0.4748	-0.354	0.723180	
centroid	NA	NA	NA	NA	
meanfun	-6.3078	0.6674	-9.452	< 2e-16	***
minfun	1.7143	0.4976	3.445	0.000571	***
maxfun	0.1786	0.7806	0.229	0.818987	
meandom	0.1320	0.5295	0.249	0.803079	
mindom	-0.7497	1.3520	-0.555	0.579211	
maxdom	-0.5286	0.5824	-0.908	0.364052	
dfrange	NA	NA	NA	NA	
modindx	-0.1777	0.7696	-0.231	0.817379	

답) 위와 같은 결과를 확인 할 수 있었는데, 유의한 변수의 개수는 총 다섯 개였습니다. 위에서 언급한바와 같이 주파수의 형태를 파악하는데 있어 유용한 평균 주파수는 당연히 목소리의 형태를 파악하는데 있어 유의한 영향을 제공할 것이라고 생각하였습니다. 첫번째 분위 수 같은 경우는 사실 세번째 분위 수와 다를 바가 없는 것 같은데, 이 변수만 유효하다는 식의 결과가 나와 많이 의아했습니다. 다음으로 스펙트럼 엔트로피는 사실 빛의 파장 분야에서 많이 쓰이는 실험방식이라 사실 소리의 파장을 분석하는데 있어서 큰 영향을 줄 것 같지는 않다는 생각을 하였습니다. 그러나 빛이든 소리가든 파장은 동일한 성질을 가진다면 유효할 것이다 라는 두가지 생각이 같이 들고 있습니다. 기본 주파수의 평균은 평균 주파수와 마찬가지로 유효하다는

생각을 하였고, 최소 기본 주파수는 왜 최소만 유효한 것인지? 최소 기본 주파수가 유효하다면 최대 기본 주파수 또한 유효해야 하는 건 아닌지? 라는 생각에 혼란스러웠습니다. 하지만, 사람의 목소리를 판가름하는 데 있어 얼마나 높은 소리를 낼 수 있느냐 보다 얼마나 낮은 목소리를 낼 수 있는가가 분류를 하기에 더 적합하다면 충분히 첫번째 분위수 주파수와 최소 우성 주파수는 유의미한 변수일 것입니다.

## 2. [Q2-2]에서 정성적으로 선택했던 변수들의 P-value를 확인하고 해당 변수가 모델링 측면에서 실제로 유효하지 않는 것인지 확인해 보시오.

답) 이전에 제가 답변한 질문에서 유효하지 않을 것이라고 판단한 변수들은 **sd(주파수 표준편차)**, **sp.ent(스펙트럼 엔트로피)**, **sfm(스펙트럼 평탄도)**, **dfrange**, **skew**, **kurt** 였습니다. 역시나 sd는 높은 p-value의 값을 가지며 유효하지 않았고, skew와 kurt 그리고 sfm 또한 유효하지 않은 변수들이 였습니다. 하지만 스펙트럼 엔트로피는 스펙트럼에 대한 저의 지식이 부족한 까닭인지 유효하다는 결과가 나왔고, dfrange는 거의 상속되다시피 하는 다른 변수가 존재하여 NA의 값으로 나왔다는 것을 확인 할 수 있었습니다.

## 3. 학습 데이터와 테스트 데이터에 대한 Confusion Matrix를 생성하고 Simple Accuracy, Balanced Correction Rate, F1-Measure를 산출하여 비교해 보시오.

답)

	lr_predicted	
lr_target	0	1
0	217	6
1	3	256

← Confusion Matrix

	TPR (Recall)	Precision	TNR	ACC
Logstic Regression	0.988417	0.9770992	0.9730942	0.9813278
	BCR	F1		
Logstic Regression	0.9807257	0.9827255		

Confusion Matrix를 보면 0범주(Negative class)는 여성을 의미하고, 1범주(Positive class)는 남성을 의미합니다. 이를 해석해보자면, 223명의 여성중에서 217명을 여성으로 제대로 예측하였고, 259명의 남성 중에서 256명의 남성을 제대로 예측했다는 것을 확인 할 수 있었습니다.

- TPR (98.8%) 는 실제 남성의 목소리를 남성의 목소리라고 제대로 예측한 확률을 의미합니다.

- TNR (97.3%) 은 실제 여성의 목소리를 여성의 목소리라고 제대로 예측한 확률을 의미



합니다.

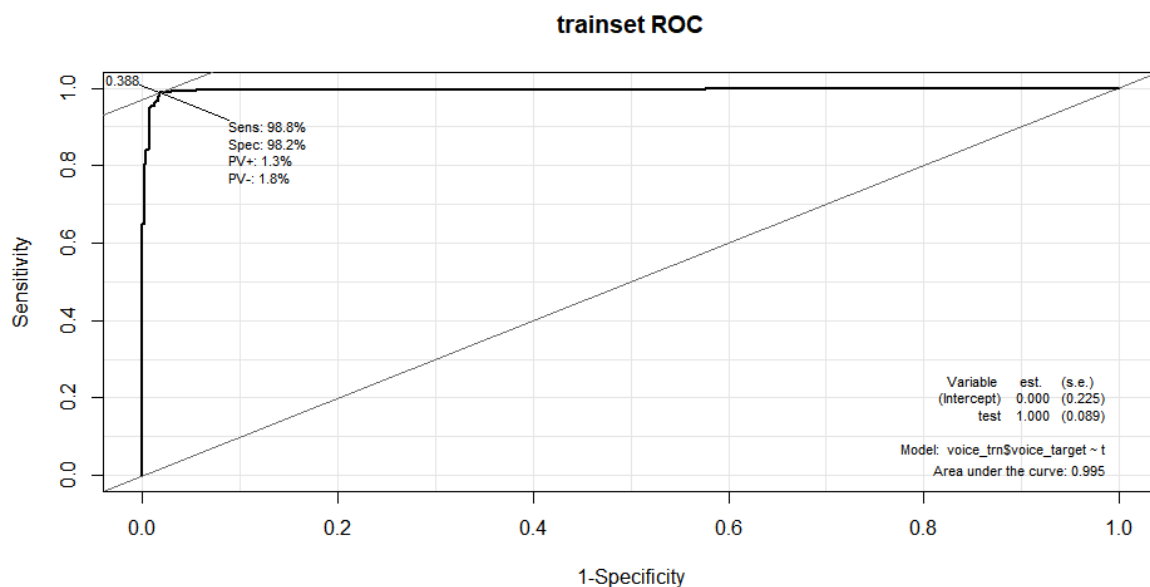
- ACC (98.1%) 는 단순 정확도를 의미하며, 실제 여성의 목소리를 여성의 목소리라고, 실제 남성의 목소리를 남성의 목소리라고 제대로 예측한 확률을 의미합니다.

- BCR (98%) 은 각 범주에 해당하는 값의 기하평균이 98%라는 것을 의미합니다.

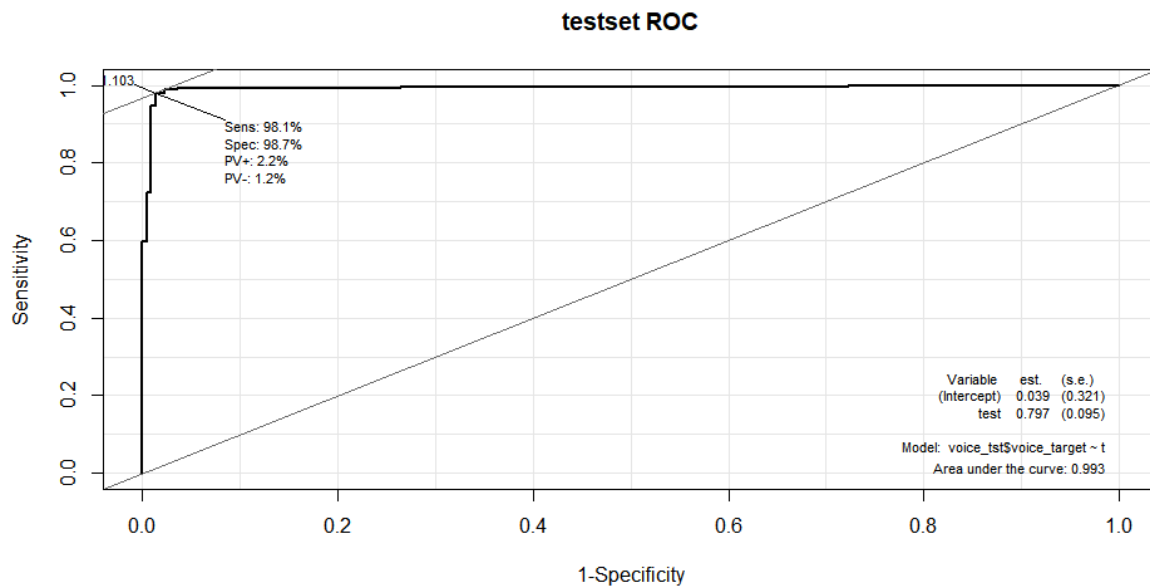
- F1-Measure (98%) 은 조화 평균을 의미합니다. 조화 평균을 사용하는 이유는 모델의 성능을 객관적으로 판단하는데 도움이 되기 위해서 입니다. 현재 98%로서 어느 한쪽으로도 치우치지 않았다는 것을 알 수 있었습니다.

#### 4. 학습 데이터와 테스트 데이터에 대한 AUROC를 산출하는 함수를 직접 작성하고 이를 사용하여 학습/테스트 데이터셋에 대한 AUROC를 비교해 보시오.

답) R-script를 보면 총 두가지의 방법으로 AUROC를 구현하였습니다. 비교를 하기 위해 사용하는 플롯은 패키지를 이용하여 만든 AUROC입니다. 우선 학습 데이터셋에 대한 AUROC를 보면 다음과 같은 결과가 나왔습니다.



Area under the curve=0.995가 나왔으며 AUROC는 그 면적이 1에 가까울수록 좋은 성능을 가진다고 평가를 하는데 매우 괜찮은 성능을 가지고 있다는 것을 확인 할 수 있었습니다. 다음으로는 테스트 데이터셋에 대한 ROC를 확인해보도록 하겠습니다.



역시나 학습데이터를 가지고 성능을 평가했을 때와 유사하게 Area under the curve는 0.993으로 아주 높은 결과가 나왔습니다. 하지만 역시 학습 데이터는 학습데이터를 학습시키고 그 결과를 통한 성능을 평가한 것이라 그런지 조금 더 높은 수치가 나온 것으로 확인할 수 있습니다.

R-script에서 AUROC를 총 두가지 방법으로 구현하여 보았는데, 두번째 방법은 반복문을 통해 ROC\_table을 채워나가는 방법으로 코드를 구현하였습니다. TPR과 FPR값 계산을 하기 더 수월하도록 lr\_response를 기준으로 내림차순 시켜주었고, lr\_response에 저장된 벡터는 positive로 분류될 확률을 의미하였습니다. 이 방법에서는 반복문을 통해 매 iteration마다 기준선 위를 Positive로 아래를 Negative로 예측하여 값을 채우는데, FPR값은 X축에 TPR값은 Y축에 찍어 ROC를 구축하였습니다. 성능은 패키지를 이용하였을 때와 거의 비슷한 결과물이 나와 분석을 할 때는 패키지를 통해 구축한 AUROC 테이블을 이용하여 분석을 진행하였습니다.

[Q7] [Q5]에서 변수 간 상관관계를 기준으로 선택한 변수들만을 사용하여 [Q6]에서 사용한 학습/테스트 70:30분할 데이터로 Logistic Regression 모델을 학습해 보시오.

1. 유의수준 0.05에서 유효한 변수의 수는 몇 개인지 확인하고 [Q6-1]의 결과와 비교하시오.

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.9964	-0.0234	0.0026	0.0907	4.6199

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )	
(Intercept)	-0.8465	0.9377	-0.903	0.366665	
meanfreq	1.8293	1.8312	0.999	0.317831	
sd	0.3405	1.1940	0.285	0.775503	
Q75	-0.6364	1.0371	-0.614	0.539500	
IQR	3.7405	0.9793	3.820	0.000134	***
skew	3.2438	2.3629	1.373	0.169816	
sp.ent	1.6148	1.0493	1.539	0.123801	
sfm	-1.2739	0.8439	-1.510	0.131148	
mode	-0.1740	0.4490	-0.388	0.698324	
meanfun	-6.4535	0.6697	-9.636	< 2e-16	***
minfun	1.5743	0.4777	3.295	0.000983	***
maxfun	0.3026	0.7704	0.393	0.694475	
meandom	0.2515	0.5057	0.497	0.618946	
mindom	-0.9693	1.2917	-0.750	0.452981	
maxdom	-0.5866	0.5641	-1.040	0.298366	
modindx	-0.3633	0.7342	-0.495	0.620762	

---

signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

답) 사실 유효 변수에 크게 달라지는 점이 없을 거라고 생각을 했으나 생각보다 많은 점이 달라졌다는 사실에 놀랐습니다. 우선 meanfreq 변수는 유효했던 변수였는데 더 이상 유효하지 않은 변수로 변했다는 것을 확인 할 수 있었습니다. 반면, IQR 같은 경우, 이전에는 공분산성을 띄어 NA값이 나와 정확히 알 수는 없었으나, 새로운 유효 변수로서 나왔다는 것을 확인 할 수 있었습니다. 그 외에 meanfun과 minfun은 그대로였다는 점을 체크할 수 있었습니다.

2. 학습 데이터와 테스트 데이터에 대한 Confusion Matrix를 생성하고 생성하고 Simple Accuracy, Balanced Correction Rate, F1-Measure를 산출한 뒤, [Q6-3]의 결과와 비교해 보시오.

답)

	lr_predicted	
lr_target	0	1
0	217	6
1	3	256

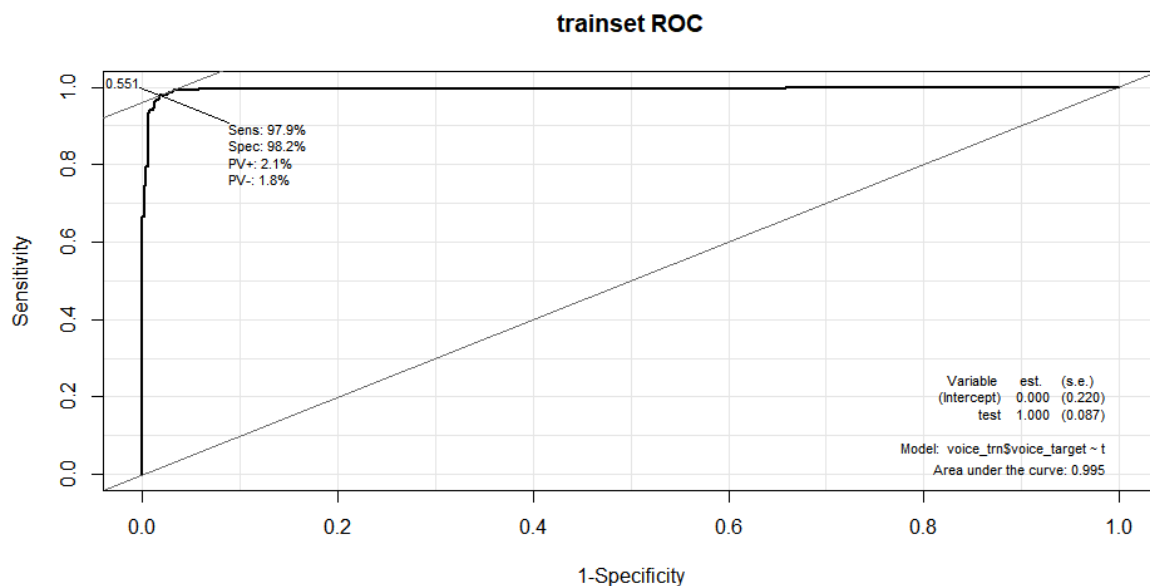
← Confusion Matrix

	TPR (Recall)	Precision	TNR	ACC	BCR	F1
Logstic Regression	0.988417	0.9770992	0.9730942	0.9813278	0.9807257	0.9827255

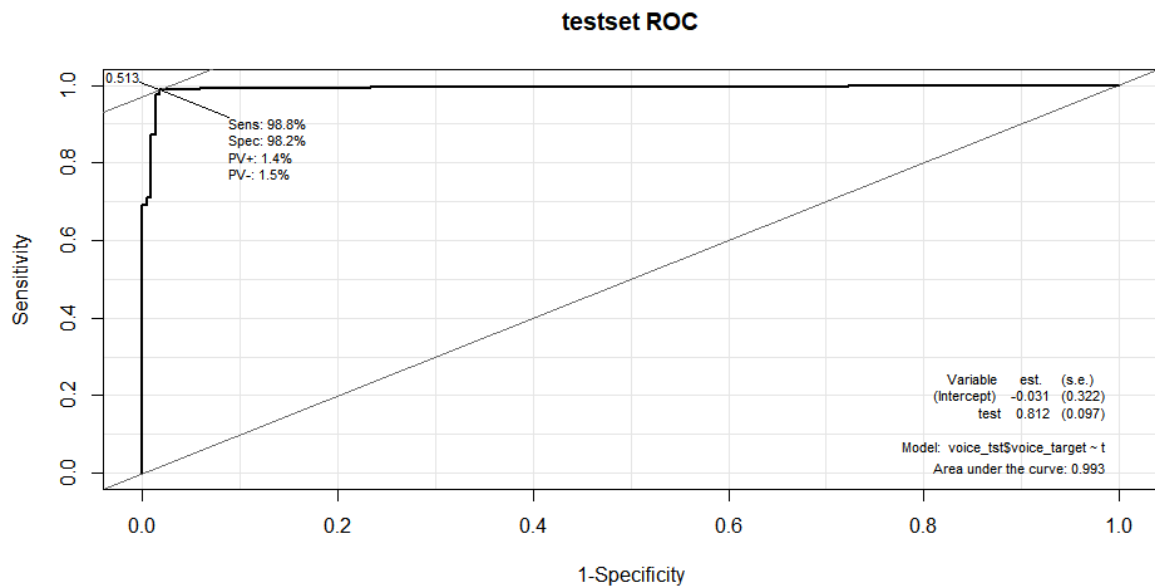
정말 의외의 결과로 값이 조금의 변동이라도 있을 줄 알았으나 6-3의 값과 완벽히 일치하는 값이 나왔다는 것을 확인 할 수 있었습니다.

3. 학습/테스트 데이터셋에 대한 AUROC를 산출하여 [Q6-4]의 결과와 비교해 보시오.

답)



7-2번 문제와는 달리 Sensitivity가 아주 미세하게 떨어졌다는 것을 확인 할 수 있었습니다. 한마디로 TPR이 아주 미세하게 떨어졌다는 것을 의미합니다.



또한, 테스트 데이터셋에 대한 ROC에서는 민감도와 특이도가 둘 다 이전보다 떨어진 모습을 보이고 있습니다. 즉, 성능이 이전에 비해 정말 미세하게 떨어진 듯 보일 수 있으나, 거의 같은 값을 가진다고 보입니다.

**[Q8] 이 외 본인이 선택한 데이터셋에 Logistic Regression을 통해 분석/검증해볼 수 있는 아이디어를 제시하고 이에 대한 절차와 분석 결과를 설명하십시오.**

답) 저는 이 문제에서 총 두가지의 궁금점을 가지고 이 문제에 임하였습니다.

```
ORtable=function(x,digits=2){
  suppressMessages(a<-confint(x))
  result=data.frame(exp(coef(x)),exp(a))
  result=round(result,digits)
  result=cbind(result,round(summary(x)$coefficient[,4],3))
  colnames(result)=c("OR","2.5%","97.5%","p")
  result
}
```

첫번째는 오즈비 였습니다. 로지스틱 회귀 모델은 오즈비야 말로 정말 로지스틱 회귀 모델의 꽃이라고 할 수 있을 것입니다. 이를 알기 위해 제가 현재 가지고 있는 데이터셋의 독립변수들의 오즈비를 구하고자 다음과 같은 코드를 돌려 다음과 같은 결과를 얻을 수 있었습니다.

```
> ORtable(full_lr)
```

	OR	2.5%	97.5%	p
(Intercept)	0.43	0.07	2.83	0.367
meanfreq	6.23	0.16	224.88	0.318
sd	1.41	0.13	13.89	0.776
Q75	0.53	0.07	4.17	0.540
IQR	42.12	6.75	322.36	0.000
skew	25.63	0.25	2788.64	0.170
sp.ent	5.03	0.69	43.28	0.124
sfm	0.28	0.05	1.38	0.131
mode	0.84	0.35	2.02	0.698
meanfun	0.00	0.00	0.01	0.000
minfun	4.83	1.96	12.88	0.001
maxfun	1.35	0.28	5.87	0.694
meandom	1.29	0.48	3.51	0.619
mindom	0.38	0.03	5.25	0.453
maxdom	0.56	0.18	1.69	0.298
modindx	0.70	0.16	2.98	0.621

오즈비를 구하는 함수를 통해 각 변수들에 대한

오즈비를 구하게 되었는데, 과연 이 테이블이 의미하는 바가 무엇인지에 대해 생각해보았습니다.

오즈비는 승산비를 의미하므로 즉 현재 데이터를 보면 하나의 예시를 들자면 meanfreq는 수치형 변수이므로 오즈비가 6.23이 나왔다는 것은 meanfreq가 1단위 증가하였을 때 오즈값이 6.23배 증가하였다는 것입니다. 하지만 현재 제가 구한 테이블을 보면 IQR같은 경우 오즈비가 42.12나 나와버리는 값을 확인할 수 있습니다. 해석을 하면 IQR이 1증가할 때 오즈값이 42.12배나 증가하였다는 것인데, 이는 IQR이 1 증가하면 1에 들어갈 확률이 저렇게 큰 배율로 증가한다는 것입니다. 사실 신뢰가 가지 않는 데이터라고 생각하였고, 어째서 이런 수치가 나올 수 있을까 분석을 해본 결과 변수가 '수치형'이라는 점에서 이상한 값이 나올 수 있다는 것을 감지하였습니다.

다음으로 제가 제 데이터셋에 대하여 궁금하였던 부분은 워낙 AUC의 값도 크게 나오고 정확도도 99%에 가까운 결과를 내다보니, 유효한 변수라고 나온 IQR, meanfun, minfun을 없앤다고 하더라도 높은 결과값이 나올 수 있을까? 만약 그럼에도 높은 결과가 나온다면 이때 유효한 변수는 무엇일까? 라는 것이었습니다. 우선 로지스틱 회귀모델을 돌려본 결과는 다음과 같았습니다.

```
glm(formula = voice_target ~ ., family = binomial, data = voice_trn)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-3.10713	-0.31770	0.03491	0.53463	2.72251

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )	
(Intercept)	0.04399	0.39822	0.110	0.9120	
meanfreq	-4.42571	0.57125	-7.747	9.38e-15	***
sd	0.87740	0.38879	2.257	0.0240	*
Q75	2.33927	0.34741	6.733	1.66e-11	***
skew	0.44564	0.98070	0.454	0.6495	
sp.ent	3.86578	0.51644	7.485	7.13e-14	***
sfm	-4.40282	0.42870	-10.270	< 2e-16	***
mode	-0.18270	0.17588	-1.039	0.2989	
maxfun	-0.22687	0.31564	-0.719	0.4723	
meandom	0.41166	0.21617	1.904	0.0569	.
mindom	1.32206	0.56351	2.346	0.0190	*
maxdom	-0.61576	0.24291	-2.535	0.0112	*
modindx	-1.79191	0.33011	-5.428	5.69e-08	***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 1559.58 on 1124 degrees of freedom  
Residual deviance: 719.76 on 1112 degrees of freedom  
AIC: 745.76

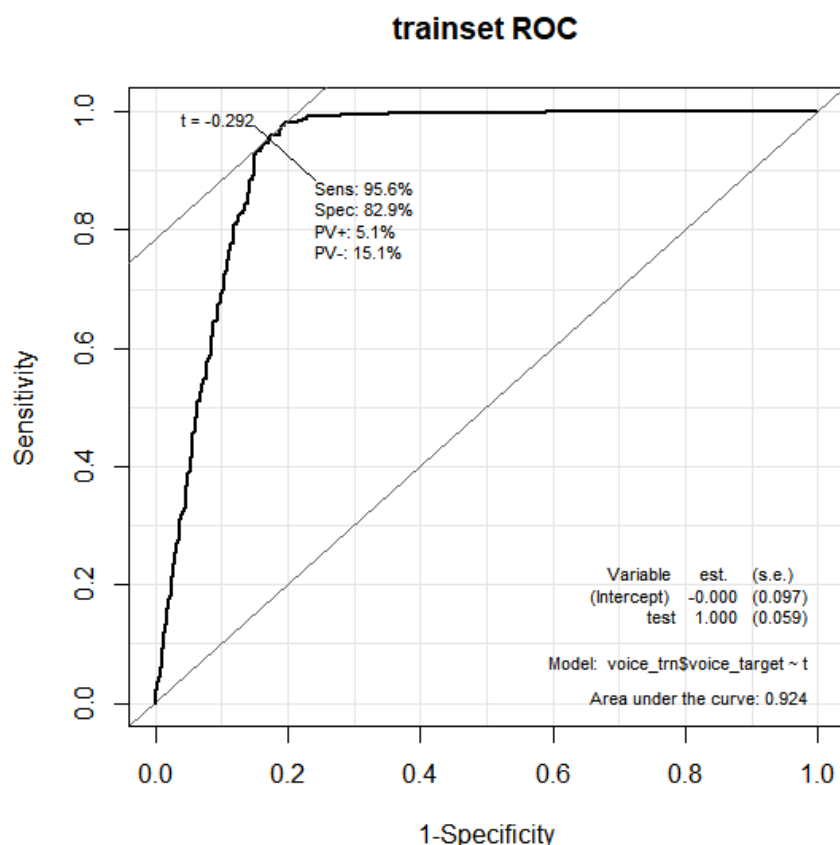
정말 이외의 결과가 나왔는데, 이 모델을 돌리기 전 저는 유효한 변수의 개수가 더 줄어들 것이

라고 생각하였습니다. 하지만 오히려 유효하지 않은 변수보다 유효한 변수가 더 많다는 결론이 나왔습니다. 이 모델의 성능을 측정해본 결과는 다음과 같습니다.

	TPR (Recall)	Precision	TNR	ACC
Logstic Regression	0.8918919	0.8716981	0.8475336	0.8713693
	BCR	F1		
Logstic Regression	0.8694299	0.8816794		

	lr_predicted	
lr_target	0	1
0	189	34
1	28	231

하지만 역시 이전 모델에서 유효하다고 나온 변수들을 빼고 돌린 탓인지, 혼동행렬을 확인해볼 결과 정확도는 더 떨어졌다는 것을 알 수 있었고, AUC 를 그려본 결과는 다음과 같습니다.



비록, AUC는 0.924로 이전에 비해 낮은 값이 나왔으나 역시나 높은 결과값을 도출해낼 수 있었습니다. 이 분석을 통해 제가 깨달을 수 있었던 점은 이 데이터셋은 정말 로지스틱 회귀분석을 하기에 잘 짜여진 데이터라는 것이었습니다. 그 이유는, 위의 과정에서 저는 유효하지 않다고 생각

되는 변수들 중에서 상관관계가 높은 변수들을 하나로 축약을 시키는 등의 행위를 하여 결과값이 크게 변하지 않았다는 것을 확인 할 수 있었습니다. 물론, 상식적으로 생각해보았을 때, 서로 비슷한 성향을 가진 변수들을 하나로 줄인다고 해서 결과에 크게 영향을 미치지 않을 것이라는 것은 쉽게 예측이 가능한 부분이었으나, 유효하다고 나온 변수들을 전부 제거해버리고 모델을 돌리게 된다면 결과에 따라 나머지 변수들이 어떤 효과를 가지고 있는지에 대해 알 수 있을 것이라고 생각하였습니다. 지금 제가 분석한 이 데이터를 자세히 보면, IQR이나 meanfun 같은 변수들은 독립변수임에도 이외의 다른 변수들을 통해 유추가 가능한 변수인 것 같습니다. 즉 독립변수지만 서로 그물망처럼 엮여 있어 충분히 높은 예측률을 보여줄 수 있었던 것 같습니다. 사실 이러한 경우는 로지스틱 회귀 모델 뿐 아니라, KNN, SVM classifier, DECISION TREE classifier 등 여러 분석 방법에서도 높은 결과를 보여줄 것이라는 예상을 할 수 있었습니다.