Using Logistic Regression for Gender Recognition Based on Audio

Phillip Efthimion

Most people would think that it is easy to guess a person's gender based on the sound of one's voice. There are a lot of factors that make up one's voice. We would like to see if we can use specific pieces of the one's voice, such as the frequencies, to determine gender.

The data consists of 3,168 recordings of voices, both male and female. Each voice was analyzed through R using the warbleR, seewave, and tuneR packages which measured 21 properties about each voice (Becker). They were analyzed only at frequencies between 0 and 280 hertz. This is because this is the human vocal range (Becker). It is unknown if subjects read from a set passage or if there was any other type of control on what was said while recordings were taken. This dataset of voices is a superset of 4 datasets: "The Harvard-Haskins Database of Regularly-Timed Speech", "Telecommunications & Signal Processing Laboratory (TSP) Speech Database at McGill University", "VoxForge Speech Corpus", and "Festvox CMU_ARTIC Speech Database at Carnegie Mellon University.

The objective of this analysis is to determine how the odds that a voice belongs to a male changes, and which variables increase the odds of the voice belonging to a male by the highest factor. We also will compare the logistic regression model with and without Principal Components Analysis using results gathered from "Using Principal Components Analysis for Gender Recognition Based on Audio" by P. Efthimion. Principal Component Analysis would be helpful to avoid multicollinearity because it reduces the number of variables in the model.

The response variable, gender, identifies one as either male or female for the purposes of this study. Males have been coded as 1 and females have been coded as 0. This coding is what allows us to perform logistic regression on the data as logistic regression is a binary

analysis. This was recorded when the subject had their voice recorded, it is not a educated guess based on other variables.

The first predictor variable is the mean frequency of each voice (meanfreq). It is measured in kilohertz (kHz). As one speaks, one does not talk at the same pitch. This variables measures the average frequency of each subject's voice.

The second predictor is the standard deviation of frequency (sd). It measures the standard deviation of each subject's voice. Standard deviation is a measure of amount of variation in a set of data.

The third predictor, median, refers to the median frequency of each voice. The median is the "middle" frequency of each voice. That is, if one was to lay out all of the frequencies in numerical order that was spoken, the median would be the one right in the middle. It is measured in kilohertz.

The fourth, fifth, and sixth predictors are the first quartile (Q25), the third quartile (Q75), and the interquartile range (IQR). The first quartile refers to the midpoint between the minimum and median numbers of the data when the data would be ranked into 4 equal groups. The third quartile refers to the midpoint between the median and the maximum. The IQR is Q75 - Q25, and is the middle 50% of the data. They are all measured in kilohertz. These are all basic statistical terms.

The seventh predictor term is skew. Skewness is a measure of the asymmetry of the probability distribution. Skewness affects a term's normality. According to R's seewave package, skewness is calculated by S = sum((x-mean(x))^3) / (N-1) / sd^3.

The eighth predictor term is kurtosis (kurt). Kurtosis is a statistical measurement related to skewness. Kurtosis measures the tails of distributions and is a measure of peakness. This could also affect normal distributions. According to R's seewave package, kurtosis is calculated by K = sum((x-mean(x))^4) / (N-1) / sd^4.

The ninth predictor term is spectral entropy (sp.ent). It is a measure for the complexity of the noise using the spectrum and amplitude. (Sueur and Lelouch)

The tenth predictor term refers to spectral flatness (sfm). Spectral flatness, also known as Wiener entropy, is a measurement that quantifies how noise-like as opposed to tone-like a sound is. It is measured from 0 to 1 with 0 being a pure tone and one being white noise. (Sueur and Simonis)

The eleventh predictor term is the mode frequency. The mode of the dataset refers to the value that occurs more times than any other value. This is a basic statistical measurement.

The twelfth predictor is centroid. Centroid is computed by C = sum(x*y) with y being the dependent variable gender. This is another measure for finding the “center” of the data. (Sueur and Simonis)

The thirteenth predictor refers to the peak frequency (peakf). The peak frequency is the maximum frequency that the subject vocalizes. It is measured in kilohertz (kHz).

The fourteenth, fifteenth, and sixteenth predictors are the average (meanfun), minimum (minfun), and maximum (maxfun) fundamental frequency measured across acoustic signal. Fundamental frequency is the lowest frequency of a periodic waveform.

The seventeenth, eighteenth, and nineteenth predictors refers to the average (meandom), minimum (mindom), and maximum (maxdom) dominant frequency measured across acoustic signal. The dominant frequency is the frequency of the wave with the highest amplitudes (Becker). People speak with multiple inflections and each inflection has different frequency peaks, the dominant sound wave, that is the one with the highest peaks, is the one we use to measure the dominant frequency. These are measured in kilohertz.

The twentieth predictor refers to the range of the dominant frequency measured across acoustic signal (dfrange). It is the range, therefore it is calculated by dfrange = maxdom - mindom. (Sueur and Simonis)

The twenty-first predictor refers to the modulation index (modindx). It is calculated as the sum of the absolute differences between adjacent measurements of fundamental frequencies divided by the frequency range (Sueur and Simonis).

Since we will be performing logistic regression, we will create histograms for all of the predictor variables in order make each variable normally distributed. We will use log, square root, inverse, and any other transformations in order to make each as normal as possible. Based upon the histogram of each of the predictor variables, we are going to take the log transformation of skewness, the minimum dominant frequency measurement, and the modulation index. We are going to take the square root transformation of spectral entropy, average dominant frequency measurement, and the range of dominant frequency. We will perform an inverse transformation on the interquartile range and kurtosis. Additionally, we will be performing the inverse of the log transformation on the maximum fundamental frequency. The inverse is taken on all of these values in order to transform them so that they have a normal distribution. Also, because some of the values for the modulation index, which are being log transformed, equal 0, we will being adding 1 to each of the values before the transformation. This needs to be done because log(0) does not equal a finite number. Besides a normal distribution, we also must check the assumptions that all values have uniform error variances and are linear.

We will now perform the logistic regression. We will use the generalized linear model function, glm, in R making sure to include that it is a binomial and to use the logit.

```
glm(voice2$gender ~ ., data = voice_b, family = binomial(link = 'logit'))
```

```
Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    -29.89759   20.33838  -1.470  0.14156
meanfreq        -7.13228   47.08454  -0.151  0.87960
sd              44.83206   35.06181   1.279  0.20102
median          -3.45556   13.39588  -0.258  0.79644
Q25            -32.44353   15.32175  -2.117  0.03422 *
Q75             35.86201   21.28060   1.685  0.09195 .
inv_IQR         -0.11117    0.05523  -2.013  0.04413 *
log_skew        -2.03234    1.23500  -1.646  0.09984 .
inv_kurt        -5.70635    3.02550  -1.886  0.05928 .
sqrt_sp.ent     60.31711   20.68014   2.917  0.00354 **
mode             3.52728    2.22449   1.586  0.11282
sfm            -12.11245    2.58291  -4.689 2.74e-06 ***
meanfun        -161.72659    8.97617 -18.017  < 2e-16 ***
minfun          31.04120   10.07710   3.080  0.00207 **
inv_log_maxfun   0.78114    1.43191   0.546  0.58539
sqrt_meandom    -0.08087    0.98299  -0.082  0.93443
log_mindom       0.68000    0.48690   1.397  0.16253
sqrt_maxdom    -20.18989   10.27395  -1.965  0.04940 *
sqrt_dfrange    20.34473   10.17848   1.999  0.04563 *
log_modindx     -4.89743    5.14001  -0.953  0.34069
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4391.78  on 3167  degrees of freedom
Residual deviance:  546.84  on 3148  degrees of freedom
AIC: 586.84
```

The logistic regression shows that at 95% confidence, Q25, IQR, spectral entropy (sp.ent), spectral flatness (sfm), average fundamental frequency measured across acoustic signal (meanfun), minimum fundamental frequency measured across acoustic signs (minfun), maximum dominant frequency measured across acoustic signal (maxdom), and range of the dominant frequency (dfrange) are all statistically significant.

```
> vif(logistic2)
     meanfreq            sd        median           Q25           Q75       inv_IQR
   147.963820     17.225059     18.452739     34.753673     23.702216      6.393449
     log_skew      inv_kurt   sqrt_sp.ent          mode           sfm       meanfun
     7.818775      5.128600      9.064504      2.348185     14.402680      1.420192
      minfun inv_log_maxfun   sqrt_meandom   log_mindom   sqrt_maxdom  sqrt_dfrange
     2.029584      1.637800      7.423885      3.498704   5821.127468   5831.283119
  log_modindx
     2.483423
```

However, the variance inflation factor tells us that there is collinearity in our model so we will need to use a selection method to reduce the variance inflation factor of the variables. After,

performing backwards elimination the variables dfrange, meanfreq, sfm, and Q25 were removed. Our model became the following:

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.1126  -0.0414   0.0006   0.1234   3.9890

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     49.92065   11.18092   4.465 8.01e-06 ***
sd              12.26156   17.10042   0.717  0.47335
median          -4.22350    6.41199  -0.659  0.51010
Q75             11.66070    8.10905   1.438  0.15044
inv_IQR         -0.28260    0.03993  -7.077 1.47e-12 ***
log_skew        -3.52384    1.11055  -3.173  0.00151 **
inv_kurt        -6.01434    2.82940  -2.126  0.03353 *
sqrt_sp.ent    -26.60523   10.63231  -2.502  0.01234 *
mode             1.78612    2.08544   0.856  0.39174
meanfun       -158.77039    8.73250 -18.182  < 2e-16 ***
minfun          27.76823   10.66572   2.604  0.00923 **
inv_log_maxfun   0.30824    1.40132   0.220  0.82590
sqrt_meandom    -0.38602    0.94691  -0.408  0.68352
log_mindom      -0.05439    0.29479  -0.184  0.85363
sqrt_maxdom      0.38562    0.33913   1.137  0.25551
log_modindx     -5.71576    5.05979  -1.130  0.25863
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4391.78  on 3167  degrees of freedom
Residual deviance:  588.67  on 3152  degrees of freedom
AIC: 620.67
```

We can now see that the statistically significant variables are the IQR, skew, kurtosis, spectral entropy, and the average and minimum fundamental frequencies measured across acoustic signal. Also we now have no variables with a high variance inflation factor (all VIF less than 10). The one cause for concern is that the AIC has risen. Since the Z-statistics for all of our statistically significant values are less than 0.05, we reject the null hypothesis.

We will now perform a Wald test, which is based on approximating MLEs. (Slides 12.7 Inferences for Logistic regression). This chi-squared test tells us that with a very low p-value of 0.0, our variables are further confirmed to be statistically significant.

```
Wald test:
----------

Chi-squared test:
X2 = 464.3, df = 15, P(> X2) = 0.0
```

```
> exp(cbind(OR = coef(logistic2d), confint(logistic2d)))
Waiting for profiling to be done...
                        OR        2.5 %       97.5 %
(Intercept)    4.789218e+21 1.760421e+12 2.138098e+31
sd             2.114103e+05 3.688375e-10 5.635865e+19
median         1.464723e-02 4.542143e-08 3.973119e+03
Q75            1.159246e+05 1.565565e-02 1.073573e+12
inv_IQR        7.538230e-01 6.949500e-01 8.129295e-01
log_skew       2.948604e-02 3.467946e-03 2.705014e-01
inv_kurt       2.443456e-03 9.763803e-06 6.476383e-01
sqrt_sp.ent    2.789303e-12 2.425863e-21 3.496019e-03
mode           5.966229e+00 9.948604e-02 3.580087e+02
meanfun        1.114031e-69 1.661720e-77 1.305297e-62
minfun         1.147067e+12 7.210168e+02 4.638216e+20
inv_log_maxfun 1.361024e+00 9.421871e-02 2.286468e+01
sqrt_meandom   6.797578e-01 1.066210e-01 4.384012e+00
log_mindom     9.470662e-01 5.302185e-01 1.686958e+00
sqrt_maxdom    1.470527e+00 7.573727e-01 2.867151e+00
log_modindx    3.293641e-03 1.902457e-07 7.734080e+01
```

These results tells us that the higher the minimum fundamental frequency, the more the odds increase that the subject is a male. The other variables with the highest increase in odds frequency per unit are the standard deviation and Q75. This is interesting as one would think that since a man's voice typically deepens in puberty that the lower the minimum fundamental frequency, the higher the odds the subject is a man, but this is not the case according to our data. That along with the Q75 suggest the opposite. However, this is only true about the fundamental frequency, it is not the case for the minimum of the dominant frequency, which has a much lower effect on the odds ratio. The variables that if increased one unit have the lowest effect on the odds that a subject is a male are the mean of the fundamental frequency and the spectral entropy whose effect are negligible.

The equation for the odds that a subject is male is: 4.79E21 + 2.11E5(sd) + 1.46E2(median) + 1.16E5(Q75) + 7.53E-1(IQR) + 2.94E-2(skew) + 2.44E3(kurtosis) + 2.78E-12(spectral entropy) + 5.96(mode) + 1.11E-69(mean of fundamental frequency) + 1.14E12(min. fund. freq.) + 1.36 (max. fund. freq.) + 6.79E-1 (avg. dom. freq.) + 9.47E-1 (min. dom. freq.) + 1.47 (max. dom. freq.) + 3.29E3 (modulation index) .

Now we will compare these results to if we had used Principal Components Analysis on our data first. Our analysis says that we need 5 components. The first component consisting of the mean of the fundamental frequency and Q25 will be our 2nd quartile component because the second quartile is every point between Q25 and mean.

The second component, consisting of IQR, skew, kurt, and spectral entropy, deals with normality. Skewness and kurtosis measure normality and tail lengths, IQR deals with the middle 50% of data and spectral entropy deals with noise which effects the normality of a distribution. Therefore, the second component is the Normal component.

The third component deals with skew, the maximum of the fundamental frequency, the minimum of the dominant frequency, and the modular index. The modulation index is calculated with absolute differences and frequency ranges. Therefore, the third component is the Range component.

Component 4 consists of the median, Q75, maximum of the dominant frequency, df range, and the modulation index. This is the Tail component because it deals with an area where outliers may occur, although the median is not affected by outliers. The 5th component deals with the minimum, maximum, and average of the fundamental frequency along with the modulation index. This will be the Fundamental Frequency component.

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.0629  -0.2254   0.0011   0.3867   3.7371

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.43729    0.07296  -5.993 2.06e-09 ***
c1          -0.90450    0.04071 -22.217  < 2e-16 ***
c2          -1.36195    0.05675 -24.001  < 2e-16 ***
c3           0.42152    0.06057   6.960 3.41e-12 ***
c4          -0.76056    0.05747 -13.233  < 2e-16 ***
c5          -2.71850    0.12152 -22.370  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4391.8  on 3167  degrees of freedom
Residual deviance: 1593.9  on 3162  degrees of freedom
AIC: 1605.9

Number of Fisher Scoring iterations: 7
```

```
> exp(cbind(OR = coef(pclm_logistic), confint(pclm_logistic)))
Waiting for profiling to be done...
                    OR      2.5 %    97.5 %
(Intercept) 0.64578557 0.55850731 0.7436139
c1          0.40474229 0.37270240 0.4372417
c2          0.25616050 0.22847381 0.2854329
c3          1.52428000 1.35604659 1.7196622
c4          0.46740678 0.41685368 0.5222655
c5          0.06597364 0.05162295 0.0831472
```

Since the Z-statistics all are less than 0.05 we reject the null hypothesis. The logistic regression on these principal components tells us that the Range component is the component that increases the odds of a subject being male the most, by a factor of 0.42 for each unit increase. Like our last model, all of the other factors also increase the odds of the subject being male for each unit increase.

In comparing our original previous model to this model using PCA, we see that the AIC has risen from 620.67 to 1605.9. One wants a model's AIC to be as low as possible so for our model using PCA to have risen by roughly 1,000 means that our model without using PCA is a better model. This seems counterintuitive since usually less dimensions in a model are preferred. What is interesting about the Range component and the highest coefficients of our previous model is none of our variables that had the highest coefficients in the previous model are in the Range component of our PCA model. This is something that should be studied in further analysis. Perhaps by accounting for more than 80% of the variation.

The previous paper concluded that more variables were statistically significant than our model using logistic regression. This could be due to following more proper procedure this time as the response variable is binary. Future analysis would look into differences between the components, though in both analyses, had 5 components, all statistically significant.

Therefore, our final model will be the one created without using principal components analysis as it is the most accurate in terms of AIC: 4.79E21 + 2.11E5(sd) + 1.46E2(median) + 1.16E5(Q75) + 7.53E-1(IQR) + 2.94E-2(skew) + 2.44E3(kurtosis) + 2.78E-12(spectral entropy) + 5.96(mode) + 1.11E-69(mean of fundamental frequency) + 1.14E12(min. fund. freq.) + 1.36 (max. fund. freq.) + 6.79E-1 (avg. dom. freq.) + 9.47E-1 (min. dom. freq.) + 1.47 (max. dom. freq.) + 3.29E3 (modulation index) .

Further analysis that I would like to do would be training the model using machine learning techniques and the "caret" package in R. This will allow us to train our model we

performed principal component analysis on and test how accurate the model is. We also will have to really look into why our logistic model was less accurate when principal component analysis was performed. I would also look into why the data says that the higher the frequency the more likely the subject is male as that seems counterintuitive though it is possible that it is not the fundamental frequency, but the dominant frequency where lower voice are socially more associated with males.

Bibliography

Becker, Kory. “GenderRecognition by Voice | Kaggle.” *Gender Recognition by Voice | Kaggle*. N.p., n.d. Web.

Becker, Kory. "Identifying the Gender of a Voice using Machine Learning." Primary Objects. N.p., 14 Jan. 2017. Web.

Hosmer, D. & Lemeshow, S. (2000). Applied Logistic Regression (Second Edition). New York: John Wiley & Sons, Inc.

Sueur, Jerome, and Caroline Simonis. “R: Spectral Properties.” *R: Spectral Properties*. N.p., n.d. Web.

Sueur, Jerome, and Laurent Lelouch. “R: Shannon and Renyi Spectral Entropy.” *R: Shannon and Renyi Spectral Entropy*. N.p., n.d. Web.

“Wiener Entropy.” *Wiener Entropy - Sound Analysis Pro*. N.p., n.d. Web.

```
# PCA on Voice Recognition

# Phillip Efthimion


# Input Data


voice <- read.csv("/Users/Phillip/Downloads/voice.csv", sep=",", header = TRUE)

names(voice)

head(voice$label)

str(voice)




##################

# Transformation#

##################


# meanfreq. No transformation improves normality


hist(voice$meanfreq)

voice$log_meanfreq <- log10(voice$meanfreq)

voice$sqrt_meanfreq <- (voice$meanfreq)^0.5

voice$inv_meanfreq <- 1 / (voice$meanfreq)

hist(voice$log_meanfreq)

hist(voice$sqrt_meanfreq)

hist(voice$inv_meanfreq)
```

```
# sd. No transformation improves normality

hist(voice$sd)
voice$log_sd <- log10(voice$sd)
voice$sqrt_sd <- (voice$sd)^0.5
voice$inv_sd <- 1 / (voice$sd)
hist(voice$log_sd)
hist(voice$sqrt_sd)
hist(voice$inv_sd)

# median. No transformation improves normality

hist(voice$median)
voice$log_median <- log10(voice$median)
voice$sqrt_median <- (voice$median)^0.5
voice$inv_median <- 1 / (voice$median)
hist(voice$log_median)
hist(voice$sqrt_median)
hist(voice$inv_median)

# Q25. No transformation improves normality

hist(voice$Q25)
voice$log_Q25 <- log10(voice$Q25)
```

```
voice$sqrt_Q25 <- (voice$Q25)^0.5

voice$inv_Q25 <- 1 / (voice$Q25)

hist(voice$log_Q25)

hist(voice$sqrt_Q25)

hist(voice$inv_Q25)


# Q75. No transformation improves normality. None look normal.


hist(voice$Q75)

voice$log_Q75 <- log10(voice$Q75)

voice$sqrt_Q75 <- (voice$Q75)^0.5

voice$inv_Q75 <- 1 / (voice$Q75)

hist(voice$log_Q75)

hist(voice$sqrt_Q75)

hist(voice$inv_Q75)


# IQR. Inverse transformation helps improve normality


hist(voice$IQR)

voice$log_IQR <- log10(voice$IQR)

voice$sqrt_IQR <- (voice$IQR)^0.5

voice$inv_IQR <- 1 / (voice$IQR)

hist(voice$log_IQR)

hist(voice$sqrt_IQR)

hist(voice$inv_IQR)
```

```
# skew. Log transformation helps improve normality.

hist(voice$skew)

voice$log_skew <- log10(voice$skew)

voice$sqrt_skew <- (voice$skew)^0.5

voice$inv_skew <- 1 / (voice$skew)

hist(voice$log_skew)

hist(voice$sqrt_skew)

hist(voice$inv_skew)

# kurt. Inverse transformation helps improve normality.

hist(voice$kurt)

voice$log_kurt <- log10(voice$kurt)

voice$sqrt_kurt <- (voice$kurt)^0.5

voice$inv_kurt <- 1 / (voice$kurt)

hist(voice$log_kurt)

hist(voice$sqrt_kurt)

hist(voice$inv_kurt)

# sp.ent. SQRT transformation helps improve normality

hist(voice$sp.ent)

voice$log_sp.ent <- log10(voice$sp.ent)
```

```
voice$sqrt_sp.ent <- (voice$sp.ent)^0.5

voice$inv_sp.ent <- 1 / (voice$sp.ent)

hist(voice$log_sp.ent)

hist(voice$sqrt_sp.ent)

hist(voice$inv_sp.ent)


# Mode. No transformation improves normality.


hist(voice$mode)

voice$log_mode <- log10(voice$mode)

voice$sqrt_mode <- (voice$mode)^0.5

voice$inv_mode <- 1 / (voice$mode)

hist(voice$log_mode)

hist(voice$sqrt_mode)

hist(voice$inv_mode)


# Centroid. No transformation improves normality.


hist(voice$centroid)

voice$log_centroid <- log10(voice$centroid)

voice$sqrt_centroid <- (voice$centroid)^0.5

voice$inv_centroid <- 1 / (voice$centroid)

hist(voice$log_centroid)

hist(voice$sqrt_centroid)

hist(voice$inv_centroid)
```

```
# Meanful. o transformation improves normality.

hist(voice$meanfun)

voice$log_meanfun <- log10(voice$meanfun)

voice$sqrt_meanfun <- (voice$meanfun)^0.5

voice$inv_meanfun <- 1 / (voice$meanfun)

hist(voice$log_meanfun)

hist(voice$sqrt_meanfun)

hist(voice$inv_meanfun)

# minion. No transformation improves normality

hist(voice$minfun)

voice$log_minfun <- log10(voice$minfun)

voice$sqrt_minfun <- (voice$minfun)^0.5

voice$inv_minfun <- 1 / (voice$minfun)

hist(voice$log_minfun)

hist(voice$sqrt_minfun)

hist(voice$inv_minfun)

# max fun. invlog best transformation.

hist(voice$maxfun)

voice$log_maxfun <- log10(voice$maxfun)
```

```
voice$sqrt_maxfun <- (voice$maxfun)^0.5
voice$inv_maxfun <- 1 / (voice$maxfun)
hist(voice$log_maxfun)
hist(voice$sqrt_maxfun)
hist(voice$inv_maxfun)
voice$inv_log_maxfun <- 1 / (log10(voice$maxfun))
hist(voice$inv_log_maxfun)

# meandom. SQRT transformation best

hist(voice$meandom)
voice$log_meandom <- log10(voice$meandom)
voice$sqrt_meandom <- (voice$meandom)^0.5
voice$inv_meandom <- 1 / (voice$meandom)
hist(voice$log_meandom)
hist(voice$sqrt_meandom)
hist(voice$inv_meandom)

# mindom. log transform.

hist(voice$mindom)
voice$log_mindom <- log10(voice$mindom)
voice$sqrt_mindom <- (voice$mindom)^0.5
voice$inv_mindom <- 1 / (voice$mindom)
hist(voice$log_mindom)
```

```
hist(voice$sqrt_mindom)

hist(voice$inv_mindom)


# maxdom. SQRT transformation.


hist(voice$maxdom)

voice$log_maxdom <- log10(voice$maxdom)

voice$sqrt_maxdom <- (voice$maxdom)^0.5

voice$inv_maxdom <- 1 / (voice$maxdom)

hist(voice$log_maxdom)

hist(voice$sqrt_maxdom)

hist(voice$inv_maxdom)


# dfrange. SQRT transformation


hist(voice$dfrange)

voice$log_dfrange <- log10(voice$dfrange)

voice$sqrt_dfrange <- (voice$dfrange)^0.5

voice$inv_dfrange <- 1 / (voice$dfrange)

hist(voice$log_dfrange)

hist(voice$sqrt_dfrange)

hist(voice$inv_dfrange)


# modindx. LOG transformation.
```

```
hist(voice$modindx)


voice$modindx <- voice$modindx + 1


voice$log_modindx <- log10(voice$modindx)
voice$sqrt_modindx <- (voice$modindx)^0.5
voice$inv_modindx <- 1 / (voice$modindx)
hist(voice$log_modindx)
hist(voice$sqrt_modindx)
hist(voice$inv_modindx)


###################
###################


# Convert labels male & female to 0 & 1


voice$gender[1:1584] <- 1
voice$gender[1585:3168] <- 0




# add 1 to $modindx so we can take the log transformation
voice$modindx <- voice$modindx + 1
```

```
voice2 <- voice[, c("meanfreq", "sd", "median", "Q25", "Q75", "inv_IQR", "log_skew", "inv_kurt",
"sqrt_sp.ent", "mode", "sfm", "centroid", "meanfun", "minfun", "inv_log_maxfun",
"sqrt_meandom", "log_mindom", "sqrt_maxdom", "sqrt_dfrange", "log_modindx", "gender")]

voice_x <- voice_x[, -21]
names(voice_x)

##############
###Logistic###
##############

voice_b <- voice_x[,-12]

logistic2 <- glm(voice2$gender ~ ., data = voice_b, family = binomial(link = 'logit'))
summary(logistic2)

#########
###VIF###
#########

library(car)
vif(logistic2)

# Remove sqrt_dfrange
voice_b1 <-voice_b[,-18]
```

```
logistic2a <- glm(voice2$gender ~ ., data = voice_b1, family = binomial(link = 'logit'))
summary(logistic2a)
vif(logistic2a)

# Remove meanfreq
voice_b2 <-voice_b1[,-1]
logistic2b <- glm(voice2$gender ~ ., data = voice_b2, family = binomial(link = 'logit'))
summary(logistic2b)
vif(logistic2b)

# Remove sfm
voice_b3 <-voice_b2[,-10]
logistic2c <- glm(voice2$gender ~ ., data = voice_b3, family = binomial(link = 'logit'))
summary(logistic2c)
vif(logistic2c)

# Remove Q25
voice_b4 <-voice_b3[,-3]
logistic2d <- glm(voice2$gender ~ ., data = voice_b4, family = binomial(link = 'logit'))
summary(logistic2d)
vif(logistic2d)
```

```
pca_logistic <- princomp(voice_b, cor = T)

summary(pca_logistic, loadings = T)

c1 <- pca_logistic$scores[,1]

c2 <- pca_logistic$scores[,2]

c3 <- pca_logistic$scores[,3]

c4 <- pca_logistic$scores[,4]

c5 <- pca_logistic$scores[,5]

pclm_logistic <- glm(voice2$gender ~ c1 + c2 + c3 + c4 + c5, family = binomial(link = 'logit'))

summary(pclm_logistic)

exp(cbind(OR = coef(pclm_logistic), confint(pclm_logistic)))
```