STA 138 Project II

Xi Cheng,

Yi Zhang,

Haozhe Gu 999200555,

Jiansen Jiang

# Introduction

Determine a person's gender as male or female from their voice seems to be intuitive to a human.  However, we are not certain about how we detect the difference between a male and a female. One of the most important factors implying gender might be the pitch. Males general have lower pitch while females have higher pitch. Because the pitch can be roughly described by the fundamental frequency, so we want to investigate the relationship between the average fundamental frequency and gender of the speaker. We also want to know whether the average fundamental frequency alone classifies subjects’ gender well. The response variable is subject’s gender and the explanatory variable is the average fundamental frequency of their speech. Of course, the sound should be processed to filter out the frequency outside human vocal range (0-280 Hz).

# Materials and Methods

The dataset comes from a project of Kaggle (https://www.kaggle.com/primaryobjects

/voicegender). They decomposed more than 3000 speeches from sources like Festvox CMU\_ARCTIC Speech Database at Carnegie Mellon University using R package. This R package analyzes .wav files and produces more than twenty acoustic properties.

First we shuffle the dataset and remove variables not in our interest. The trimmed dataset has variable label with level male and level female and variable meanfun showing average fundamental frequency for each observation measured in Hz. We will model the probability of the male speaker with the Simple Logistic Regression model seen below:

We will first test the goodness of fit with Pearson residual statistics along with AIC. Then the estimated parameter will then be analyzed with Wald hypothesis test and Wald confidence interval under 95% confidence level. We then interpret odd ratio and median effective level in terms of the question. We will also discuss the accuracy of the logistic regression with true positive and false positive rate. Finally, we will determine the presence of outliers with Pearson residuals, Standardized Pearson residuals, standardized deviance residuals and likelihood residuals.