Vowel Analysis Final Report

Faith Hardie   DS303, SP25   Prof. Amber Camp

2025-03-14

Table of contents

Vowel Analysis Final Report

Load packages

```
Warning: package 'lmerTest' was built under R version 4.4.3
```

```
Warning: package 'phonR' was built under R version 4.4.3
```

Load data

Load your personal data (make sure you update from P101 -> your P#)

```
# read in data
P112 <- read_csv("data/P112 (2).csv")
```

```
Rows: 102 Columns: 26
-- Column specification --------------------------------------------------------
Delimiter: ","
chr (17): ppt, word, ipa, arpa, onset, offset, environment, real_word, sex, ...
dbl  (9): item_num, rep, f0, f1, f2, duration, age, years_uni, age_learned_en

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
# convert variables to factor where needed
convert_to_factor <- function(df, cols) {
  df[cols] <- lapply(df[cols], as.factor)
  return(df)
}

P112 <- convert_to_factor(P112, c("ppt", "word", "ipa", "arpa", "onset", "offset", "environme

# remove a couple words you won't be needing
P112 <- P112 %>%
  dplyr::filter(!word %in% c("cot", "caught")) # added dplyr to specify which 'filter' to use
```

Class data:

```
# read in data
all_data <- read_csv("data/DS303_combined.csv")
```

```
Rows: 1279 Columns: 26
-- Column specification --------------------------------------------------------
Delimiter: ","
chr (17): ppt, word, ipa, arpa, onset, offset, environment, real_word, sex, ...
dbl  (9): item_num, rep, f0, f1, f2, duration, age, years_uni, age_learned_en

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
# convert variables to factor where needed
all_data <- convert_to_factor(all_data, c("ppt", "word", "ipa", "arpa", "onset", "offset", "e

# remove a couple words you won't be needing
all_data <- all_data %>%
  dplyr::filter(!word %in% c("cot", "caught"))
```

```
##Summary and skim
#summary(P112)
skimP112 <-  skimr::skim(P112)
#view(P112)

#summary(all_data)
skimalldata <-  skimr::skim(all_data)
view(all_data)
class(all_data)
```

```
[1] "spec_tbl_df" "tbl_df"      "tbl"         "data.frame"
```

```
##Removing outliers
## remove outliers
class_clean <- all_data %>%
  group_by(ppt, ipa) %>%
  mutate(
    f1_z = (f1 - mean(f1)) / sd(f1),
    f2_z = (f2 - mean(f2)) / sd(f2)
  ) %>%
  filter(abs(f1_z) <= 1.25, abs(f2_z) <= 1.25)

P112_clean <- P112 %>%
  group_by(ppt, ipa) %>%
  mutate(
    f1_z = (f1 - mean(f1)) / sd(f1),
    f2_z = (f2 - mean(f2)) / sd(f2)
  ) %>%
  filter(abs(f1_z) <= 1.25, abs(f2_z) <= 1.25)
```

Explain the Data

(1 point)

In paragraph form:

- Describe where the data comes from
- Summarize the contents of the data (how many observations, participants, items, etc.)
- Mention any pre-processing steps taken. For example, I pre-processed this data by removing words that were obviously mispronounced before even sending it to you. Then, above, you converted certain variables to factor and removed the words “cot” and “caught”, which are not relevant to your investigation. Have you done any additional processing?

ANSWER: The data in DS303_combined.csv comes from our classes recorded voices. The process of getting this data started with completing a google form with questions regarding our personal backgrounds such as gender, languages we speak, our place of birth, and other questions based on circumstances that effect how we speak. We then recorded our voices using PRATT software and Dr. Camp organized our classes recordings into a data set.

The data for the class wide data set, DS303_combined.csv has 26 variables, 13 participants, 1201 observations and 32 itemsitems. Each student had 29 words.

In addition to Dr. Camp removing mispronounced words, we converted certain variables to factors and removed the words "cot" and "caught." Independently, I cleaned the data set by removing any outliers by filtering it based on their z core, through the process of cleaning my data, computing the z scores of f1 and f2, and removing anything outside of 1.25 standard deviation.

Variables of Interest

(1 point)

For this project, you will explore and analyze the **class-wide data set**. In paragraph form:

- Briefly introduce the purpose of this project
- Identify and explain your variables of interest
- State research questions or hypotheses about this data

The purpose of this project is to analyze the different components of how our class says words, including the vowel, onset, offset, and environmental factors that can affect how we speak. My variables of interest are real word, f1, and f2. I am interested to see if whether the word is real or not affects how the vowel is pronounced.

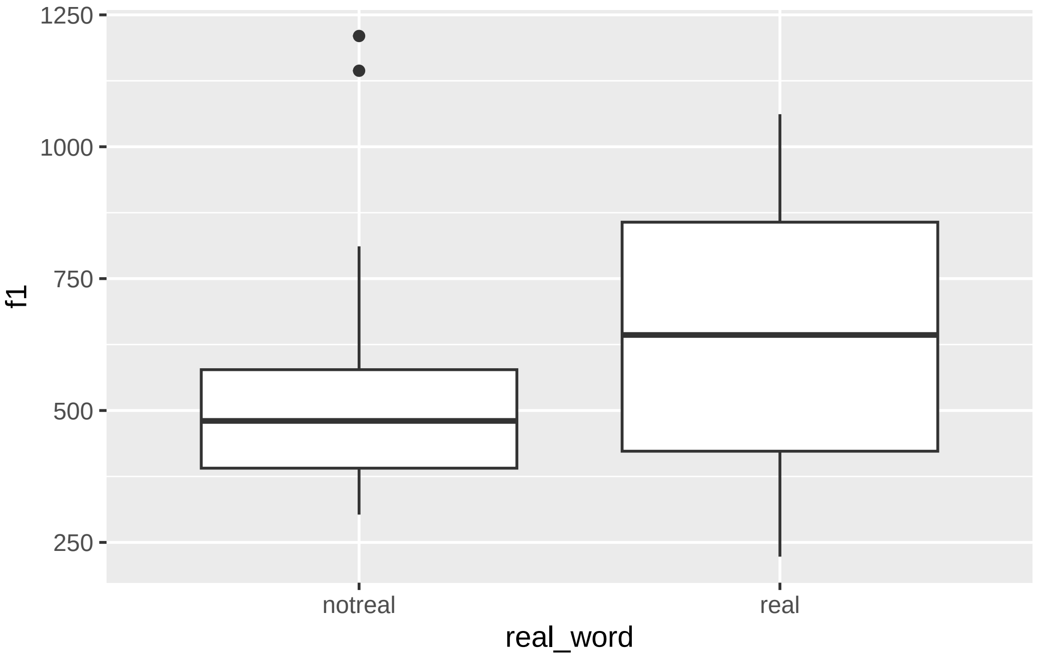
Research question: Can we predict if a word is real or not based on f1 and/or f2? vice versa.

EDA and Vowel Plots

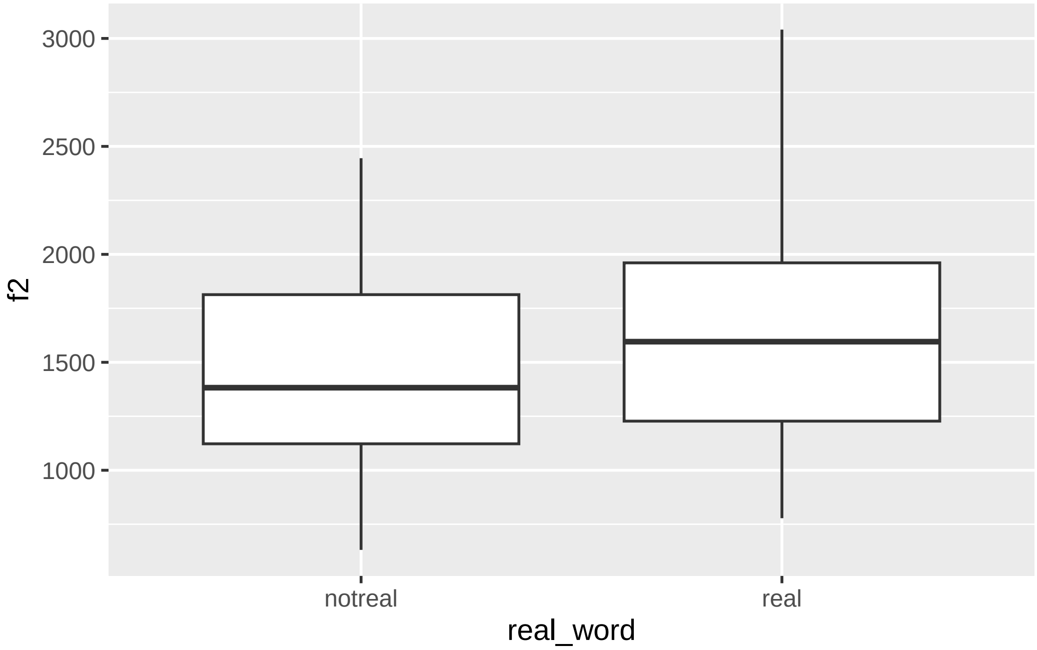
(3 points)

- Generate two vowel plots using `phonR`: one for your own data, and one for class data
- In a couple sentences, state your observations. Do you see any patterns or differences?
- Include at least one visual that supports your hypothesis/justifies your models below, and explain

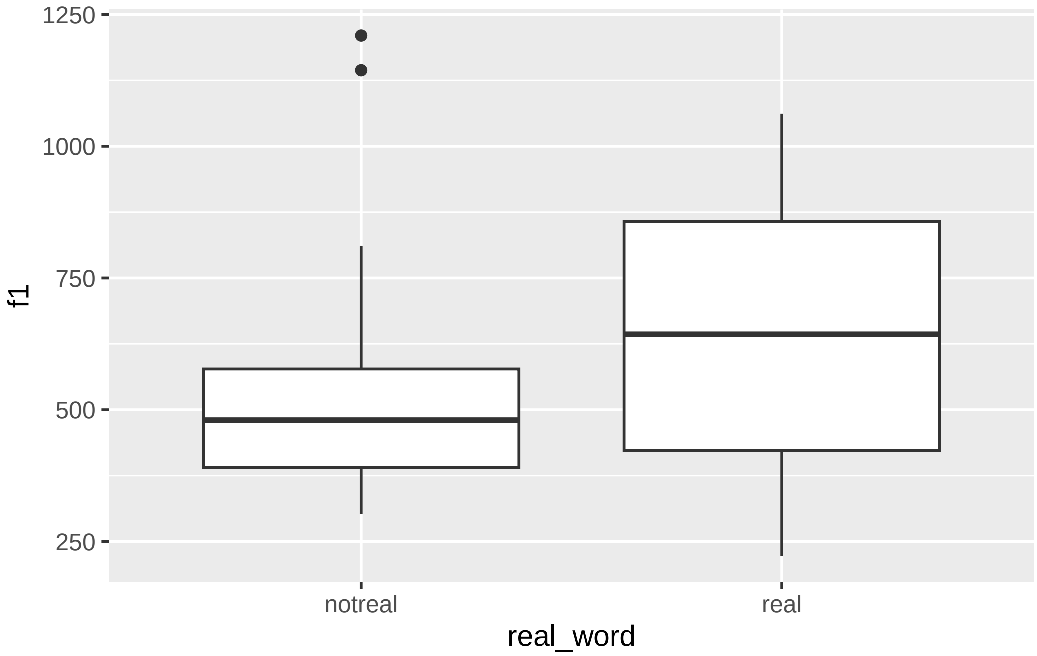
```r
ggplot(P112, aes(x = real_word, y = f1)) +
  geom_boxplot()
```



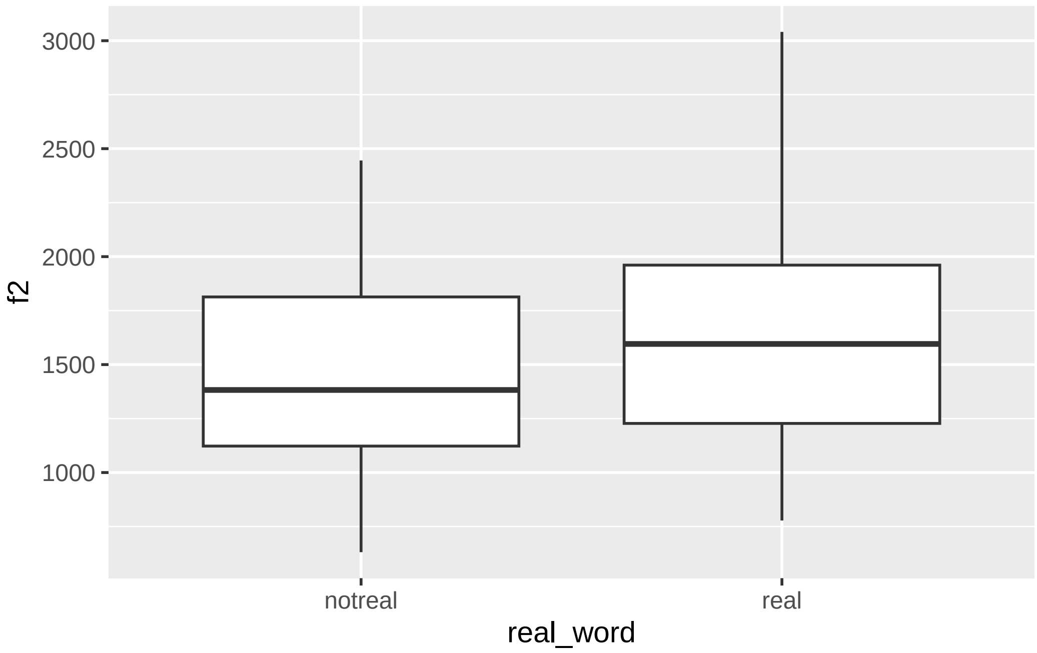
```r
ggplot(class_clean, aes(x=real_word, y=f2)) +
  geom_boxplot()
```



```
ggplot(P112, aes(x = real_word, y = f1)) +
  geom_boxplot()
```



```
ggplot(class_clean, aes(x=real_word, y=f2)) +
  geom_boxplot()
```



```
#ggplot(class_clean, aes(x= f1, y=f2, fill=ethnicity))+
#  geom_point()
par()
```

```
$xlog
[1] FALSE

$ylog
[1] FALSE

$adj
[1] 0.5

$ann
[1] TRUE

$ask
[1] FALSE
```

```
$bg
[1] "transparent"

$bty
[1] "o"

$cex
[1] 1

$cex.axis
[1] 1

$cex.lab
[1] 1

$cex.main
[1] 1.2

$cex.sub
[1] 1

$cin
[1] 0.15 0.20

$col
[1] "black"

$col.axis
[1] "black"

$col.lab
[1] "black"

$col.main
[1] "black"

$col.sub
[1] "black"

$cra
[1] 10.8 14.4

$crt
```

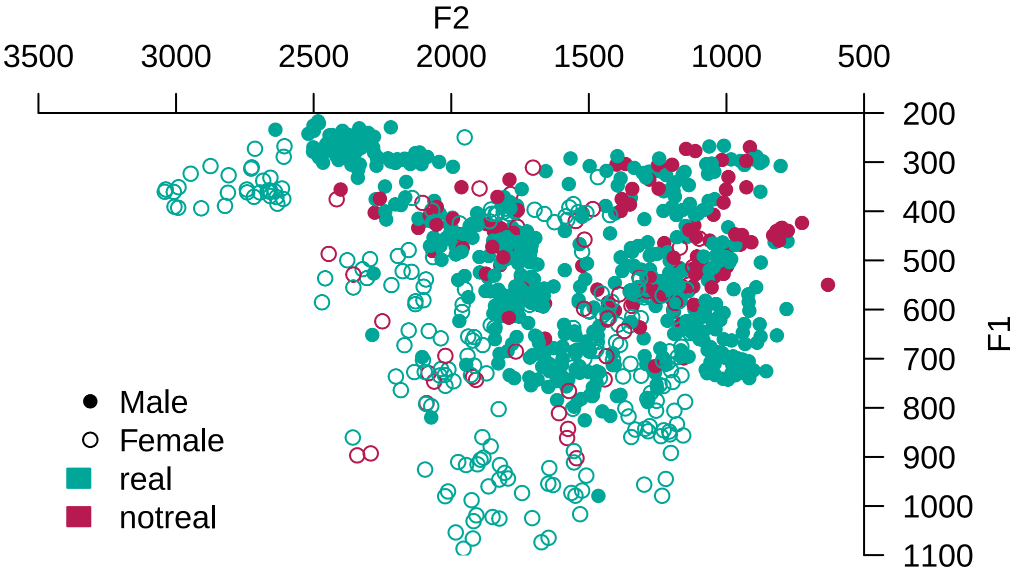
```
[1] 0

$csi
[1] 0.2

$cxy
[1] 0.03521127 0.12048193

$din
[1] 5.5 3.5

$err
[1] 0

$family
[1] ""

$fg
[1] "black"

$fig
[1] 0 1 0 1

$fin
[1] 5.5 3.5

$font
[1] 1

$font.axis
[1] 1

$font.lab
[1] 1

$font.main
[1] 2

$font.sub
[1] 1

$lab
[1] 5 5 7
```

```
$las
[1] 0

$lend
[1] "round"

$lheight
[1] 1

$ljoin
[1] "round"

$lmitre
[1] 10

$lty
[1] "solid"

$lwd
[1] 1

$mai
[1] 1.02 0.82 0.82 0.42

$mar
[1] 5.1 4.1 4.1 2.1

$mex
[1] 1

$mfcol
[1] 1 1

$mfg
[1] 1 1 1 1

$mfrow
[1] 1 1

$mgp
[1] 3 1 0
```

```
$mkh
[1] 0.001

$new
[1] FALSE

$oma
[1] 0 0 0 0

$omd
[1] 0 1 0 1

$omi
[1] 0 0 0 0

$page
[1] TRUE

$pch
[1] 1

$pin
[1] 4.26 1.66

$plt
[1] 0.1490909 0.9236364 0.2914286 0.7657143

$ps
[1] 12

$pty
[1] "m"

$smo
[1] 1

$srt
[1] 0

$tck
[1] NA

$tcl
```

```
[1] -0.5

$usr
[1] 0 1 0 1

$xaxp
[1] 0 1 5

$xaxs
[1] "r"

$xaxt
[1] "s"

$xpd
[1] FALSE

$yaxp
[1] 0 1 5

$yaxs
[1] "r"

$yaxt
[1] "s"

$ylbias
[1] 0.2
```

```
with(class_clean, plotVowels(f1, f2, var.sty.by=sex, var.col.by = real_word, pretty = TRUE,
```



Patterns, observations: For the EDA part of this assignment, I did 4 box plots using ggplot and 1 dot plot using phonr. For both my personal data and the class data, there is a small difference in both formant 1 and 2 depending on whether the word is real or not real, however, this doesn't necessary mean that the vowel formants changed because of the words realness, it could be dependent on other variables, how how the words that aren't real are commonly pronounced (the way the tongue sits when the words are said) compared to if the word is real. The difference between real and not real also doesn't seem to be statistically significant from these graphs, but I would have to run my regression models to fully understand the relationship.

Model Selection and Justification

(3 points)

- You will build and analyze **two different statistical models** to investigate the relationship between your predictors and outcome variable
- The two models should differ in some way (e.g., one could include an additional or interaction term while the other does not)
- What statistical models will you use to investigate the relationship between your predictors and outcome variable? (linear vs. logistic regression? mixed effects model?)
- Why did you select these models?

- Which variable(s) are included?

  Answer: I will be using logistic regression to investigate the relationship between my predictors and outcome variable because my predictor variable real_word has only two options, real or not_real, so it is binary, and a logistic model predicts probability. I want to know the probability that f1 can predict real_word. However, I also wanted to investigate if f1 could predict if the word is real or not, and for that I need to use linear regression, since the predictor f1 is not a binary or binomial predictor, and its values are outside of 0 or 1. The variables I am including are real_word, f1, f2, and sex.

```
m1 <- glm( real_word ~ f1 , data = class_clean, family = binomial)
summary(m1)
```

```

Call:
glm(formula = real_word ~ f1, family = binomial, data = class_clean)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) 0.7629831  0.2822942   2.703  0.00688 **
f1          0.0015183  0.0005203   2.918  0.00352 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 810.30  on 880  degrees of freedom
Residual deviance: 801.44  on 879  degrees of freedom
AIC: 805.44

Number of Fisher Scoring iterations: 4
```

```
m2 <- glm(real_word ~ f2, data = class_clean, family = binomial)
summary(m2)
```

```

Call:
glm(formula = real_word ~ f2, family = binomial, data = class_clean)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) 0.3369275  0.3163458   1.065    0.287
```

```
f2          0.0007913  0.0002023   3.911 9.19e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 810.30  on 880  degrees of freedom
Residual deviance: 793.83  on 879  degrees of freedom
AIC: 797.83

Number of Fisher Scoring iterations: 4
```

```r
m3 <- glm(real_word ~ f1*sex, data = class_clean, family = binomial)
summary(m3)
```

```
Call:
glm(formula = real_word ~ f1 * sex, family = binomial, data = class_clean)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.1966741  0.5310228   2.254   0.0242 *
f1           0.0007918  0.0008210   0.964   0.3348
sexMale     -0.6926450  0.6400952  -1.082   0.2792
f1:sexMale   0.0012693  0.0010896   1.165   0.2440
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 810.30  on 880  degrees of freedom
Residual deviance: 800.09  on 877  degrees of freedom
AIC: 808.09

Number of Fisher Scoring iterations: 4
```

```r
m4 <- glm(real_word ~ f2*sex, data = class_clean, family = binomial)
summary(m4)
```

```
Call:
```

```
glm(formula = real_word ~ f2 * sex, family = binomial, data = class_clean)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)  
(Intercept)  0.5014043  0.6614715   0.758    0.448  
f2           0.0006784  0.0003758   1.805    0.071 .
sexMale     -0.2455634  0.7595046  -0.323    0.746  
f2:sexMale   0.0001771  0.0004525   0.391    0.695  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 810.30  on 880  degrees of freedom
Residual deviance: 793.64  on 877  degrees of freedom
AIC: 801.64

Number of Fisher Scoring iterations: 4
```

```r
m5 <- lm(f1 ~ real_word, data=class_clean)
summary(m5)
```

```
Call:
lm(formula = f1 ~ real_word, data = class_clean)

Residuals:
    Min      1Q  Median      3Q     Max 
-337.32 -143.79    1.48  118.60  532.91 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)     506.91      14.59  34.753  < 2e-16 ***
real_wordreal    47.23      16.03   2.945  0.00331 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 179.8 on 879 degrees of freedom
Multiple R-squared:  0.009772,  Adjusted R-squared:  0.008645 
F-statistic: 8.674 on 1 and 879 DF,  p-value: 0.003312
```

```r
class_clean$sex <- relevel(class_clean$sex, ref = "Male")
m6 <- lm(f1 ~ real_word *sex, data=class_clean)
summary(m6)
```

```
Call:
lm(formula = f1 ~ real_word * sex, data = class_clean)

Residuals:
    Min      1Q  Median      3Q     Max 
-392.72 -122.27    4.03  125.94  465.36 

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)    
(Intercept)               468.17      16.18  28.932  < 2e-16 ***
real_wordreal              45.93      17.87   2.570   0.0103 *  
sexFemale                 140.21      30.78   4.555 5.99e-06 ***
real_wordreal:sexFemale   -12.21      33.64  -0.363   0.7168    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 169.7 on 877 degrees of freedom
Multiple R-squared:   0.12, Adjusted R-squared:  0.117 
F-statistic: 39.87 on 3 and 877 DF,  p-value: < 2.2e-16
```

```r
m7 <- lm(f1 ~ real_word *sex, data=class_clean)
summary(m7)
```

```
Call:
lm(formula = f1 ~ real_word * sex, data = class_clean)

Residuals:
    Min      1Q  Median      3Q     Max 
-392.72 -122.27    4.03  125.94  465.36 

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)    
(Intercept)               468.17      16.18  28.932  < 2e-16 ***
real_wordreal              45.93      17.87   2.570   0.0103 *  
sexFemale                 140.21      30.78   4.555 5.99e-06 ***
```

```
real_wordreal:sexFemale   -12.21      33.64  -0.363   0.7168
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 169.7 on 877 degrees of freedom
Multiple R-squared:   0.12, Adjusted R-squared:  0.117
F-statistic: 39.87 on 3 and 877 DF,  p-value: < 2.2e-16
```

Model Comparisons and Best Fit

(3 points)

- Build and run both models and display their summaries
- Compare the two models, assess model fit, and determine the better fitting one

```
anova(m1, m3)
```

```
Analysis of Deviance Table

Model 1: real_word ~ f1
Model 2: real_word ~ f1 * sex
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       879     801.44
2       877     800.09  2   1.3508    0.509
```

```
#anova(m2, m4)
#anova(m5, m6)
```

Interpretation of Results

(3 points)

- Interpret coefficients and significance
- Explain how the predictor variable(s) influence the outcome

Answer: The p-value 0.509 is not statistically significant, so adding the variable sex in m3 wasn't necessarily an improvement. While the deviance (801.44-800.09) reduced, it was small, so also not statistically significant. In m1, f1 was the only predictor for real_word, versus in m3, both f1 and the sex where my predictors. In both m1 and m3, my intercepts were positive, meaning that f1 and real_word have a positive relationship. However in m3, the coefficient for male is negative, because it is being compared to female. According to my models, f1 is a predictor of real_word, and sex with f1 is not a predictor of whether a word is real or not.

Discussion and Conclusion

(3 points)

- Summarize key findings
- Discuss implications
- Mention limitations

Though testing multiple models, I can conclude that the research question: "can f1 predict if a word is real or not" can be answered based on these models 1 and 3, yes. However, that may not be because the word is real or not, but rather because that just how the vowel is pronounced. I do not think you could apply this relationship to other data outside of this data set. I also realized, that of course sex is not a predictor of real_word, because male or female doesn't change if the word is real or not. There can however be differences in how they pronounce those words, and there are. In my EDA analysis dotplot, females (open circles) were much higher on the plot than males. Overall, these regression models are limited to this data set.