Lab 09

Due April 15 at 11:59 PM

Group 2-8: Dav King, Vivian Zhang, Reesey Du Pont, Eesha Yaqub

```r
library(tidyverse)
library(tidymodels)
library(knitr)
library(viridis)
```

```r
parkinsons = read_csv("parkinsons.csv")
```

Exercise 1

```r
park = parkinsons %>%
  mutate(name = str_remove_all(name, "_[^_]+$"))
```

```r
park %>%
  filter(status == 0) %>%
  select(name) %>%
  unique() %>%
  kable()
```

name
phon_R01_S07
phon_R01_S10
phon_R01_S13
phon_R01_S17
phon_R01_S42
phon_R01_S43
phon_R01_S49
phon_R01_S50

Exercise 2

```r
test <- park %>%
  group_by(status) %>%
  arrange(name) %>%
  slice(1:24) %>%
  ungroup()
nrow(test)
```

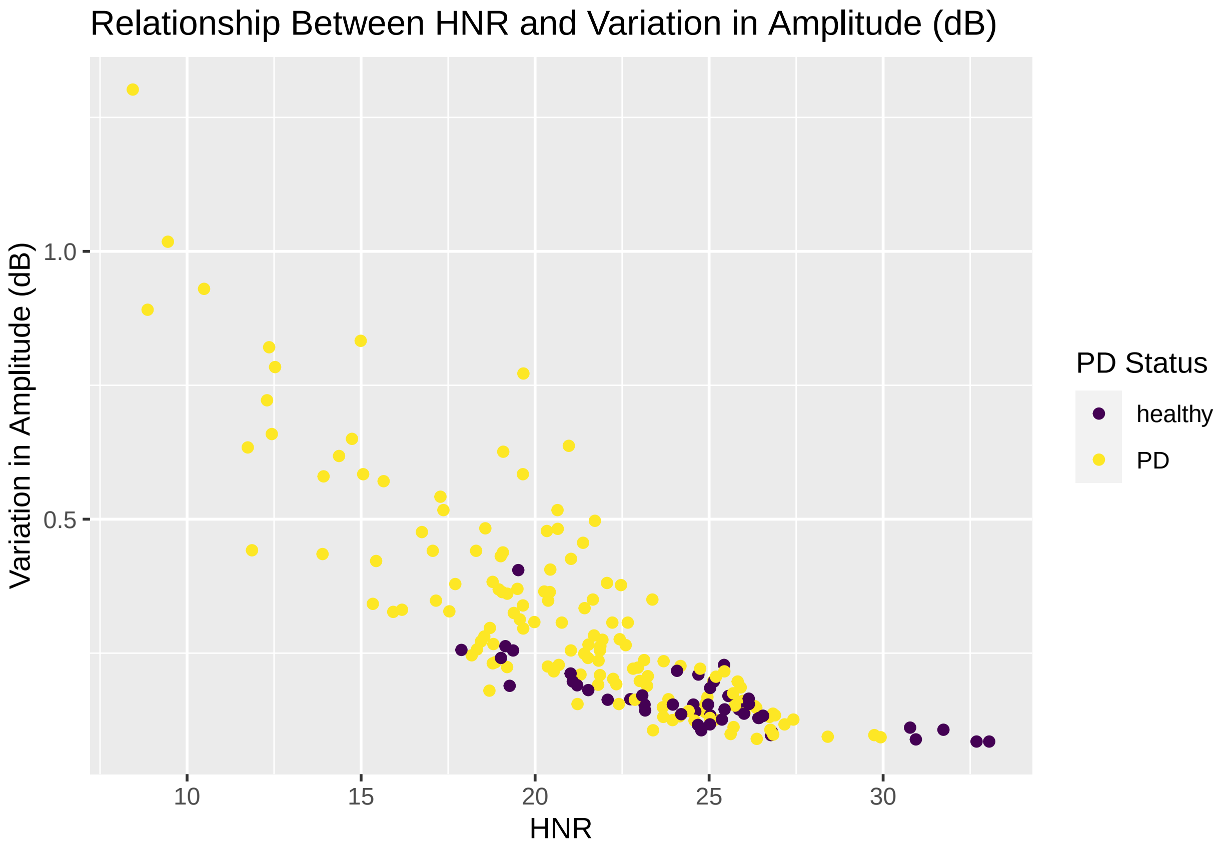
```
## [1] 48
```

```
train <- park %>%
  group_by(status) %>%
  arrange(name) %>%
  slice(-c(1:24)) %>%
  ungroup()
nrow(train)
```

```
## [1] 147
```

Exercise 3

```
park %>%
  mutate(PD_status = ifelse(status == 0, "healthy", "PD")) %>%
  ggplot(mapping = aes(x = HNR, y = shimmer, color = PD_status)) +
  geom_point() +
  labs(x = "HNR", y = "Variation in Amplitude (dB)",
       title = "Relationship Between HNR and Variation in Amplitude (dB)") +
  scale_color_viridis(discrete = TRUE, option = "D", name = "PD Status")
```



We observe a non-linear, negative correlation overall between HNR and amplitude variance. More specifically, those with PD had a larger range in amplitude variance, while the amplitude variance of healthy individuals was more clustered at lower values. Similarly, no healthy participant had an HNR lower than 17.5, while participants with PD had HNR values getting even lower than 10.

Exercise 4

```
fit_1 <- logistic_reg() %>%
  set_engine("glm") %>%
  fit(as.factor(status) ~ HNR + PPE + jitter + shimmer,
      data = train, family = "binomial")
tidy(fit_1)
```

```
## # A tibble: 5 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)  -12.2       4.42     -2.75  0.00595
## 2 HNR            0.328     0.140     2.35  0.0190
## 3 PPE           21.0       6.33      3.31  0.000921
## 4 jitter      -104.      160.       -0.651 0.515
## 5 shimmer       14.8       5.91      2.50  0.0123
```

There are three predictors that are significant at the alpha = 0.05 level: HNR, PPE, and shimmer. All three of these predictors are significant because their p-value levels are less than alpha = 0.05.

Exercise 5

```
prediction = predict(fit_1, test, type = "prob")
test_result = test %>%
  mutate(predicted_prob_pd = prediction$.pred_1) %>%
  mutate(prob_50 = if_else(predicted_prob_pd > 0.5, 1, 0)) %>%
  mutate(prob_75 = if_else(predicted_prob_pd > .75, 1, 0)) %>%
  mutate(prob_90 = if_else(predicted_prob_pd > .9, 1, 0))
test_result %>%
  count(status, prob_50)
```

```
## # A tibble: 3 x 3
##   status prob_50     n
##    <dbl>   <dbl> <int>
## 1      0       0    10
## 2      0       1    14
## 3      1       1    24
```

```
test_result %>%
  count(status, prob_75)
```

```
## # A tibble: 4 x 3
##   status prob_75     n
##    <dbl>   <dbl> <int>
## 1      0       0    22
## 2      0       1     2
## 3      1       0     1
## 4      1       1    23
```

```
test_result %>%
  count(status, prob_90)
```

```
## # A tibble: 3 x 3
##   status prob_90     n
##    <dbl>   <dbl> <int>
## 1      0       0    24
## 2      1       0     7
## 3      1       1    17
```

With a decision boundary of 50%, our model yields 14 false positives and 0 false negatives. With a decision boundary of 75%, our model yields 2 false positives and 1 false negative. With a decision boundary of 90%, our model yields 0 false positives and 7 false negatives.

Using this model as a diagnostic tool for PD, we would want to use the 75% boundary. Though it was the only level to have both false positives and false negatives, it also had the fewest incorrect measures overall - thus, it is the most accurate threshold.