Exercise8

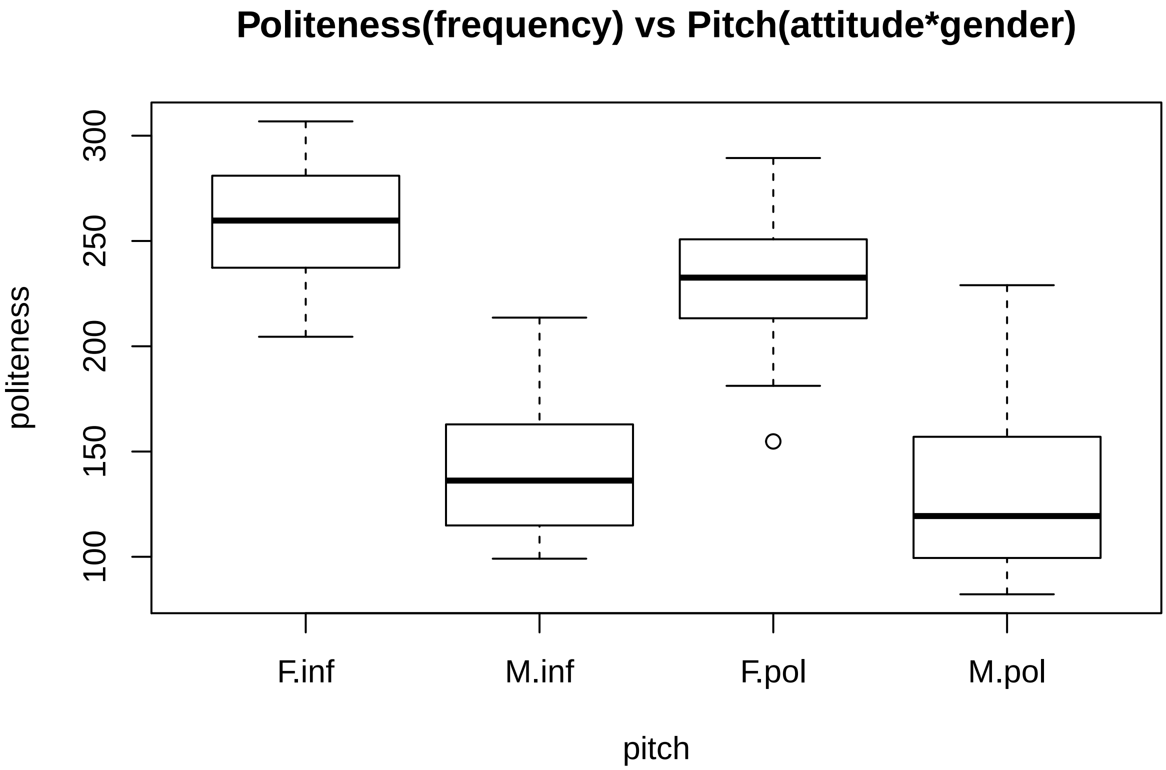
Mohammad Imtiaz Nur

4/2/2021

ID: 1878074

1.0:

```r
data <- read.csv("politeness_data.csv", header = TRUE)
boxplot(frequency~gender*attitude, data = data,
        xlab='pitch',
        ylab='politeness',
        main="Politeness(frequency) vs Pitch(attitude*gender)")
```



In the above plot, for both informal or polite attitudes, male pitching range is lower than female pitching range. Also, t is quite clear that there is a difference between female polite and informal frequency. Though the difference in frequency is very low for male, but polite frequency is slightly lower than male informal frequncy.

1.1:

```
model_1 <- lm(frequency~attitude, data = data)
summary(model_1)
```

```
##
## Call:
## lm(formula = frequency ~ attitude, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -103.488  -62.122    9.044   51.178  105.044
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   202.59      10.08  20.107   <2e-16 ***
## attitudepol   -18.23      14.34  -1.272    0.207
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 65.3 on 81 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.01958,    Adjusted R-squared:  0.007475
## F-statistic: 1.618 on 1 and 81 DF,  p-value: 0.2071
```

As the p value (0.207) is greater than the significance level (0.05), only attitude variable alone can’t make the model significant. Also the R-squared value(0.01958) signifies that our model takes only 1.958% data from the entire dataset which makes the model irrelevant.

1.2:

```
model_2 <- lm(frequency~gender, data = data)
summary(model_2)
```

```
##
## Call:
## lm(formula = frequency ~ gender, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -92.186 -28.426  -2.676  23.124  90.124
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  246.986      5.680   43.48   <2e-16 ***
## genderM     -108.110      8.081  -13.38   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 36.81 on 81 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.6884, Adjusted R-squared:  0.6846
## F-statistic:   179 on 1 and 81 DF,  p-value: < 2.2e-16
```

As the p value (< 2.2e-16) is very much lower than the significance level (0.05), the gender variable appears to build a very significant. model. Also the R-squared value(0.6884) signifies that our model takes 68.84% data from the entire dataset which makes the model a good fit for analysis.

1.3:

```
model_3 <- lm(frequency~attitude*gender, data = data)
summary(model_3)
```

```
##
## Call:
## lm(formula = frequency ~ attitude * gender, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -78.486 -27.383  -0.986  20.570  96.020
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          260.686      7.784  33.491   <2e-16 ***
## attitudepol          -27.400     11.008  -2.489   0.0149 *
## genderM             -116.195     11.008 -10.556   <2e-16 ***
## attitudepol:genderM   15.890     15.664   1.014   0.3135
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 35.67 on 79 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.7147, Adjusted R-squared:  0.7038
## F-statistic: 65.95 on 3 and 79 DF,  p-value: < 2.2e-16
```

As the p value (< 2.2e-16) is very much lower than the significance level (0.05), the attitude*gender variables together appears to build a very significant. model. Also the R-squared value(0.7038) signifies that our model takes 70.38% data from the entire dataset which makes the model a good fit for analysis.

1.4:

ANOVA test to compare the significance of all three models.

```
anova(model_1,model_2)
```

```
## Analysis of Variance Table
##
## Model 1: frequency ~ attitude
## Model 2: frequency ~ gender
##   Res.Df    RSS Df Sum of Sq F Pr(>F)
## 1     81 345341
## 2     81 109751  0    235590
```

```
anova(model_2,model_3)
```

```
## Analysis of Variance Table
##
## Model 1: frequency ~ gender
## Model 2: frequency ~ attitude * gender
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     81 109751
## 2     79 100511  2    9240.2 3.6313 0.03099 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(model_1,model_3)
```

```
## Analysis of Variance Table
##
## Model 1: frequency ~ attitude
## Model 2: frequency ~ attitude * gender
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     81 345341
## 2     79 100511  2    244830 96.216 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the ANOVA test, we can summarize that model_3 can be used for more accurate prediction than the other two models.