STA302 Proposal

Bilin

2023-10-05

load the library

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.2     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.2     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(knitr)
library(dplyr)
# set wd
# setwd("~/Desktop/Github Projects/STA302")
```

extract the subset of data we need

```
coffee_df <- read.csv(file = "merged_data_cleaned.csv") %>%
  rename("Ratings" = Total.Cup.Points) %>%
  mutate(.before = 1, InSpecies = ifelse(Species == "Arabica", 1, 0)) %>%
  select(Species, InSpecies, Aroma, Flavor, Aftertaste, Acidity, Sweetness, Ratings)
```

numerical summary

```
summary(coffee_df)
```

```
##    Species            InSpecies          Aroma           Flavor
##  Length:1339        Min.   :0.0000   Min.   :0.000   Min.   :0.00
##  Class :character   1st Qu.:1.0000   1st Qu.:7.420   1st Qu.:7.33
##  Mode  :character   Median :1.0000   Median :7.580   Median :7.58
##                     Mean   :0.9791   Mean   :7.567   Mean   :7.52
##                     3rd Qu.:1.0000   3rd Qu.:7.750   3rd Qu.:7.75
##                     Max.   :1.0000   Max.   :8.750   Max.   :8.83
##    Aftertaste       Acidity        Sweetness         Ratings
##  Min.   :0.000   Min.   :0.000   Min.   : 0.000   Min.   : 0.00
##  1st Qu.:7.250   1st Qu.:7.330   1st Qu.:10.000   1st Qu.:81.08
##  Median :7.420   Median :7.580   Median :10.000   Median :82.50
##  Mean   :7.401   Mean   :7.536   Mean   : 9.857   Mean   :82.09
##  3rd Qu.:7.580   3rd Qu.:7.750   3rd Qu.:10.000   3rd Qu.:83.67
##  Max.   :8.670   Max.   :8.750   Max.   :10.000   Max.   :90.58
```

```
summary_table <- coffee_df %>%
  summarise(
    Aroma_mean = mean(Aroma),
    Flavor_mean = mean(Flavor),
    Aftertaste_mean = mean(Aftertaste),
    Sweetness_mean = mean(Sweetness),
    Acidity_mean = mean(Acidity),
    Aroma_sd = sd(Aroma),
    Flavor_sd = sd(Flavor),
    Aftertaste_sd = sd(Aftertaste),
    Sweetness_sd = sd(Sweetness),
    Acidity_sd = sd(Acidity),
    Aroma_max = max(Aroma),
    Flavor_max = max(Flavor),
    Aftertaste_max = max(Aftertaste),
    Sweetness_max = max(Sweetness),
    Acidity_max = max(Acidity),
    Aroma_min = min(Aroma),
    Flavor_min = min(Flavor),
    Aftertaste_min = min(Aftertaste),
    Sweetness_min = min(Sweetness),
    Acidity_min = min(Acidity),
)

summary_frame = data.frame(
  Variables = c("Flavor", "Aroma", "Sweetness", "Aftertaste", "Acidity"),
  Min = c(summary_table$Flavor_min, summary_table$Aroma_min, summary_table$Sweetness_min, summary_table$
  Max = c(summary_table$Flavor_max, summary_table$Aroma_max, summary_table$Sweetness_max, summary_table$
  Mean = c(summary_table$Flavor_mean, summary_table$Aroma_mean, summary_table$Sweetness_mean, summary_ta
  SD = c(summary_table$Flavor_sd, summary_table$Aroma_sd, summary_table$Sweetness_sd, summary_table$Afte
)

kable(summary_frame, format = "markdown", caption = "Coffee Ratings Dataset Numerical Summary",
      col.names = c("Variable Name", "Minimum", "Maximum", "Mean", "Standard Deviation"),
      align = "c", longtable = TRUE, digits = 3)
```

Table 1: Coffee Ratings Dataset Numerical Summary

Variable Name	Minimum	Maximum	Mean	Standard Deviation
Flavor	0	8.83	7.520	0.398
Aroma	0	8.75	7.567	0.378
Sweetness	0	10.00	9.857	0.616
Aftertaste	0	8.67	7.401	0.404
Acidity	0	8.75	7.536	0.380

```
coffee_df %>% group_by(Species) %>%
  summarise(num_species = n()) %>%
  rename(`Number of Observations` = num_species) %>%
  kable(align = "c", longtable = TRUE)
```

Species	Number of Observations
Arabica	1311
Robusta	28

The coffee ratings dataset contains 1339 observations, with 1311 of these being of the "Arabica" species, while the other 28 are of the "Robusta" species. The minimum for all variables is 0, while the maximum rating is given in the Sweetness category. Sweetness also has the highest mean and standard deviation.

Fit the model

```
model <- lm(Ratings ~ InSpecies + Aroma + Flavor + Aftertaste + Acidity + Sweetness, data = coffee_df)
model
```

```
##
## Call:
## lm(formula = Ratings ~ InSpecies + Aroma + Flavor + Aftertaste +
##     Acidity + Sweetness, data = coffee_df)
##
## Coefficients:
## (Intercept)    InSpecies        Aroma       Flavor   Aftertaste      Acidity
##       7.166       -2.530        1.364        2.338        2.525        1.272
##   Sweetness
##       2.153
```

assumption checking

```
# fitted values
y_hat <- fitted(model)
# residual
e_hat <- resid(model)
```

```r
# attached these columns to coffee_df
coffee_df <- coffee_df %>%
  mutate("y_hat" = fitted(model)) %>%
  mutate("e_hat" = resid(model))


# write.csv(coffee_df, "CoffeeRatings.csv", row.names = FALSE)
```

residual-fitted value plot

```r
plot(x = y_hat, y = e_hat, main = "Residual vs Fitted", xlab = "Fitted",
     ylab = "Residuals")
# mtext(~italic("Fig.1: The residuals versus fitted values plot from the model."),
#       side = 1, line = 3, outer = TRUE, adj = 0.5)
title(sub = ~italic("Fig.1: The residuals versus fitted values plot from the model."))
```

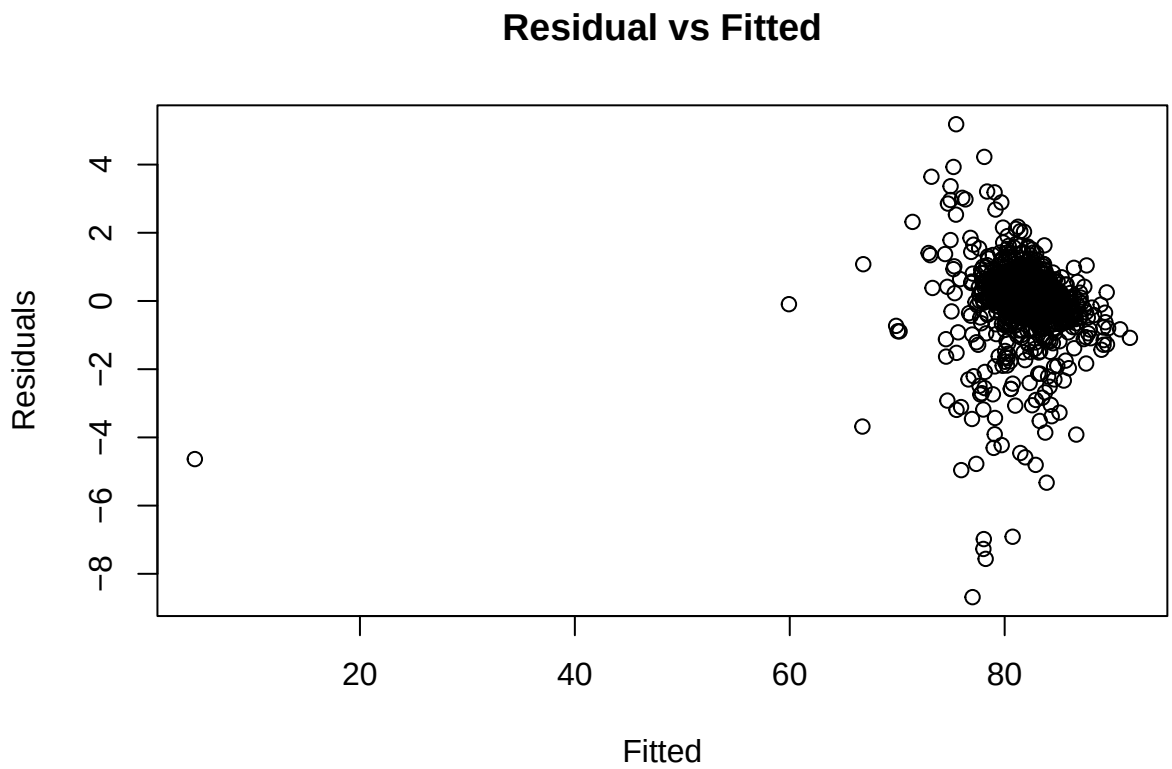

Fig.1: The residuals versus fitted values plot from the model.

```r
# plot1 <-
  # ggplot(data = coffee_df,
  #       mapping = aes(x = y_hat, y = e_hat))+
  # geom_point(alpha = 0.5, size = 2)+
  # theme_bw()+
  # theme(plot.title = element_text(hjust = 0.5))+
```

```
  # labs(x = "Fitted", y = "Residuals", title = "Residual vs. Fitted")
# ggsave(plot1, file = "ResidFitted.png", dpi = 600, width = 5, height = 5)
```

all residual vs. predictors

```
par(mfrow = c(2, 3),oma=c(4, 2, 0, 0)+0.1)

plot(x = coffee_df$InSpecies, y = e_hat,
                      main = "Residual vs. Species", xlab = "Species",
                      ylab = "Residual")
plot(x = coffee_df$Aroma, y = e_hat,
                     main = "Residual vs. Aroma", xlab = "Aroma",
                     ylab = "Residual")
plot(x = coffee_df$Flavor, y = e_hat,
                      main = "Residual vs. Flavor", xlab = "Flavor",
                      ylab = "Residual")
plot(x = coffee_df$Aftertaste, y = e_hat,
                     main = "Residual vs. Aftertaste", xlab = "Aftertaste",
                     ylab = "Residual")
plot(x = coffee_df$Acidity, y = e_hat,
                      main = "Residual vs. Acidity", xlab = "Acidity",
                      ylab = "Residual")

plot(x = coffee_df$Sweetness, y = e_hat,
                     main = "Residual vs. Sweetness", xlab = "Sweetness",
                     ylab = "Residual")
mtext(~italic("Fig.4: All residuals versus predictors plots from the model."), side = 1, line = 3, oute
```

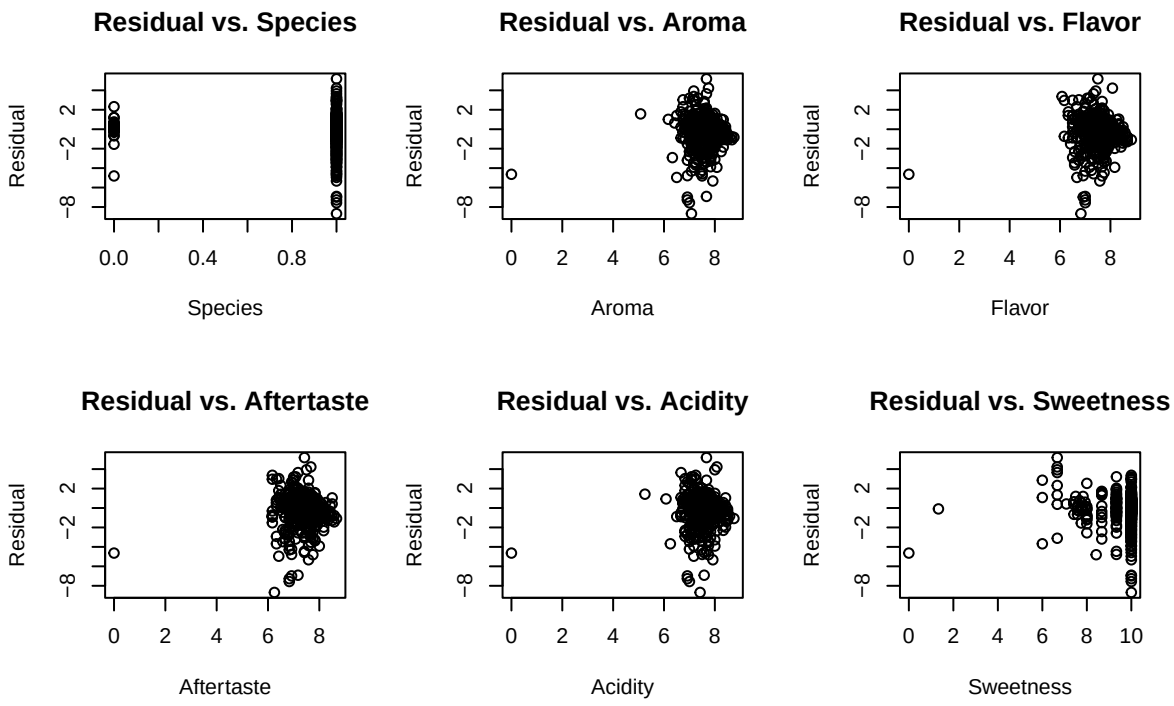

Fig.4: All residuals versus predictors plots from the model.

```
# title(sub = "Figure4: All residuals versus predictors from the model.", adj = 0.5)
```

checking normality of errors

```
qqnorm(e_hat)
qqline(e_hat)
title(sub = ~italic("Fig.3: The normal Q-Q plot of residues from the model."))
```

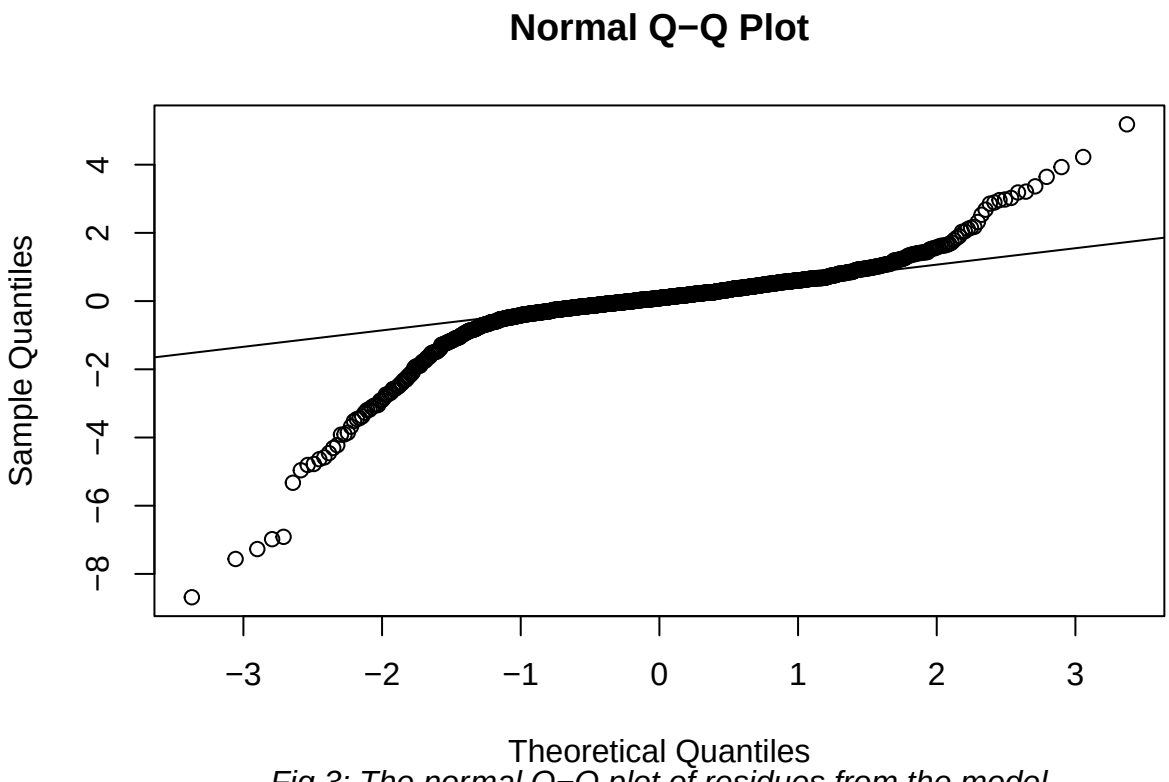

Fig.3: The normal Q−Q plot of residues from the model.

Response vs. Fitted

```
# plot3 <-
  # ggplot(data = coffee_df,
  #               mapping = aes(x = y_hat, y = Ratings))+
  # geom_point(alpha = 0.5, size = 2)+
  # theme_bw()+
  # theme(plot.title = element_text(hjust = 0.5))+
  # labs(x = "Fitted", y = "Ratings", title = "Response vs. Fitted")
# ggsave(plot3, file = "ResponseFitted.png", dpi = 600, width = 5, height = 5)
plot(x = y_hat, y = coffee_df$Ratings, main = "Response vs. Fitted", xlab = "Fitted",
     ylab = "Response")
title(sub = ~italic("Fig.2: The reponse (coffee ratings) versus fitted values plot from the model."))
```

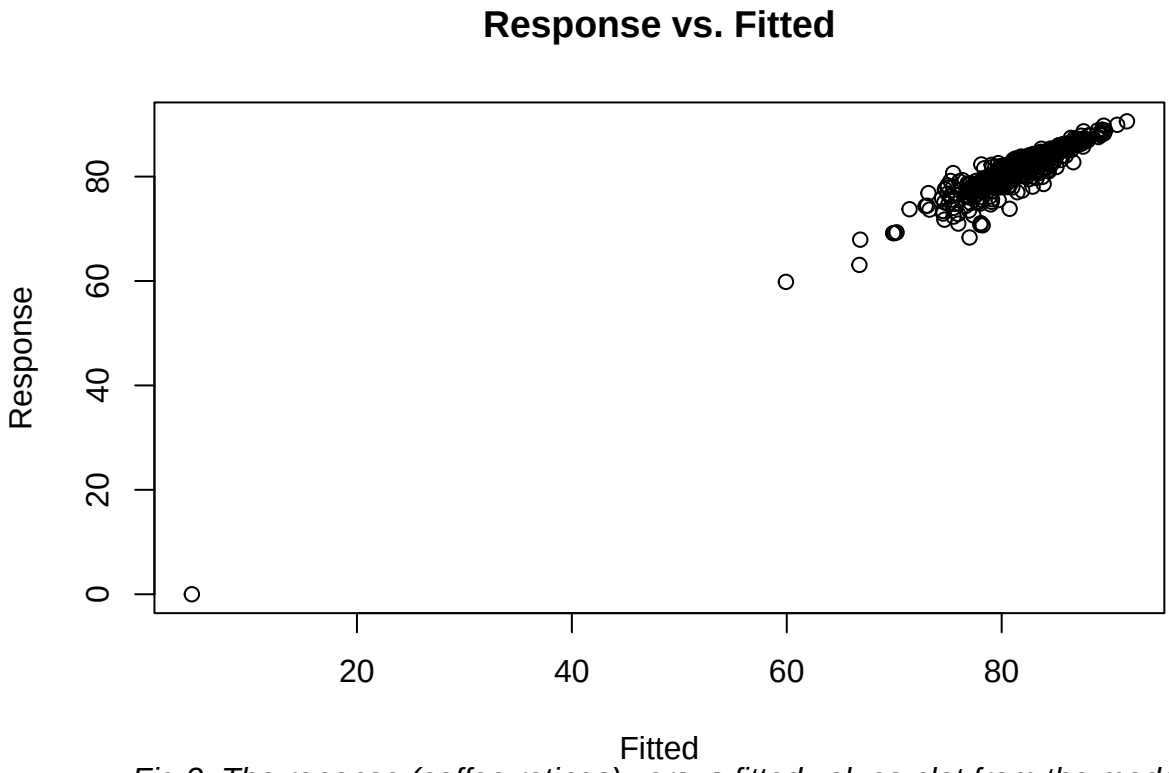

Fig.2: The reponse (coffee ratings) versus fitted values plot from the model.

pair plots of every predictors

```
pairs(coffee_df[, c(2:7)])
title(sub = ~italic("Fig.5: The pairwise scatter plots of all predictors."))
```

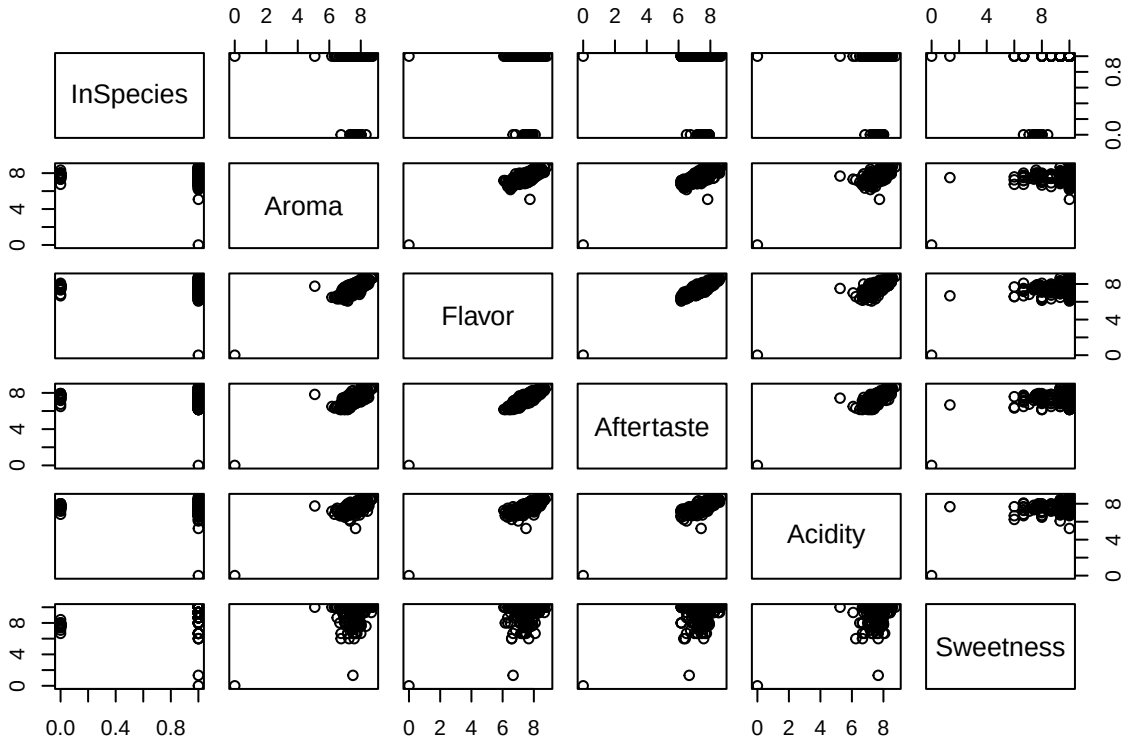

Fig.5: The pairwise scatter plots of all predictors.