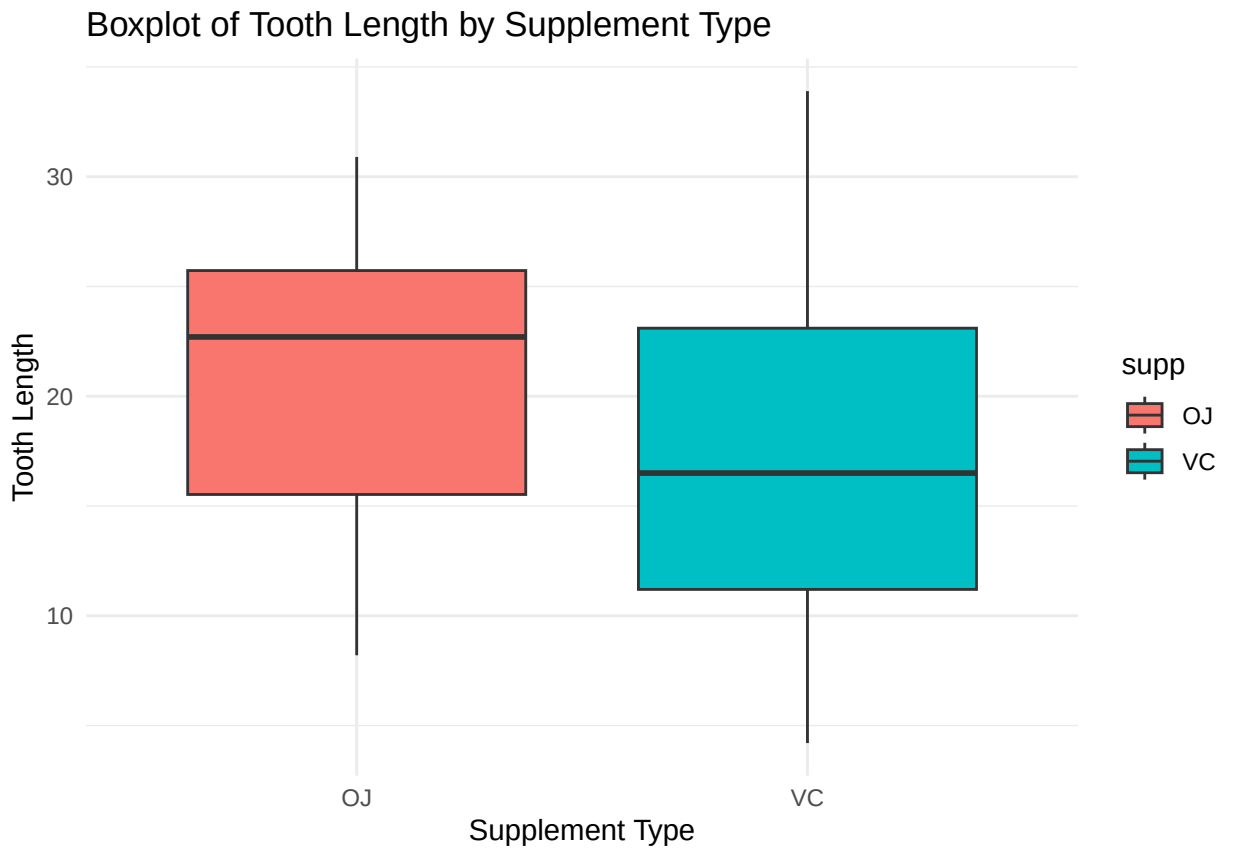
hw10

Jiayi

2024-11-17

Q1: boxplots examined the effects of different levels of supp on len

```
# Boxplot for supp
ggplot(ToothGrowth, aes(x = supp, y = len, fill = supp)) +
  geom_boxplot() +
  labs(title = "Boxplot of Tooth Length by Supplement Type", x = "Supplement Type", y = "Tooth Length") +
  theme_minimal()
```

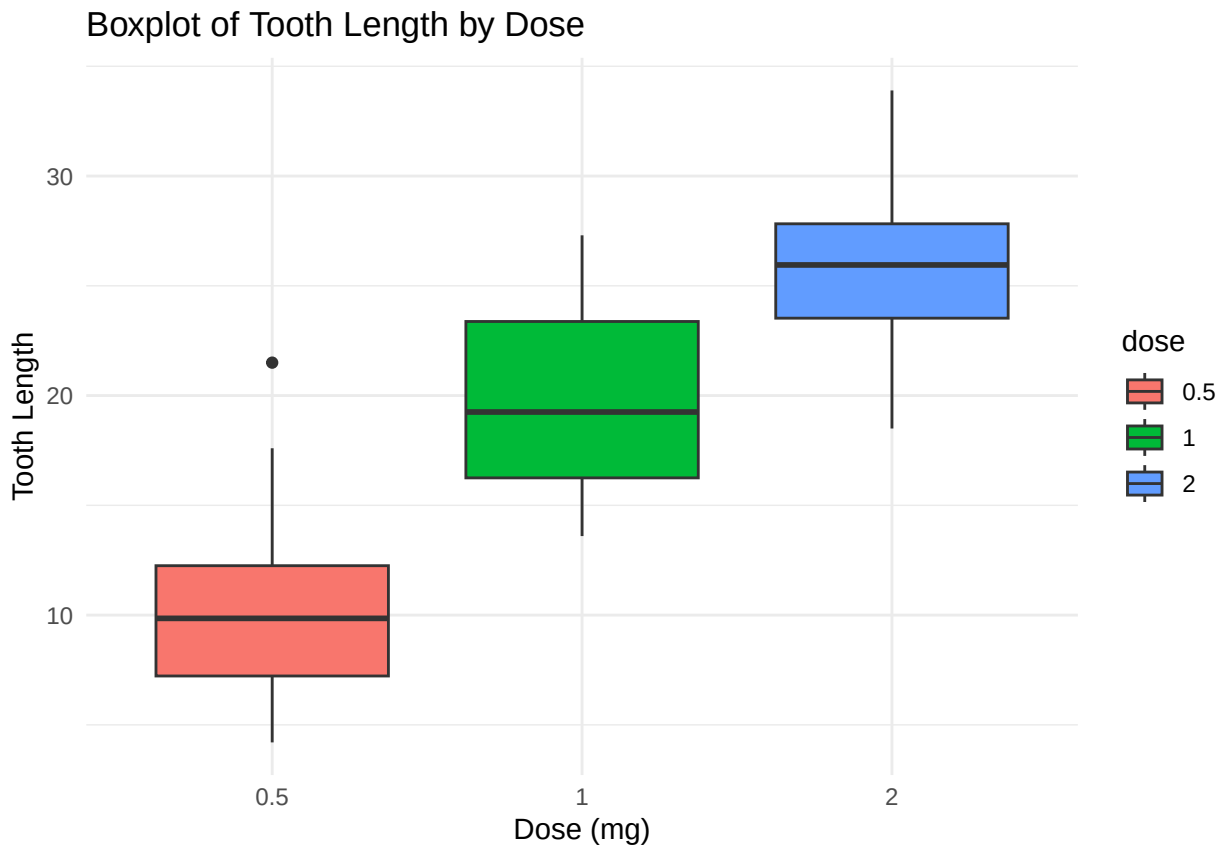


The boxplot shows that the supplement type OJ (orange juice) has a slightly higher median tooth length compared to VC (ascorbic acid). The range of tooth lengths for OJ is narrower compared to VC. This suggests that the delivery method of vitamin C (OJ vs. VC) might have an impact on tooth growth, with *OJ generally leading to greater lengths.*

boxplots examined the effects of different levels of dose on len

```r
# Convert dose to factor for boxplot
ToothGrowth$dose <- as.factor(ToothGrowth$dose)

# Boxplot for dose
ggplot(ToothGrowth, aes(x = dose, y = len, fill = dose)) +
  geom_boxplot() +
  labs(title = "Boxplot of Tooth Length by Dose", x = "Dose (mg)", y = "Tooth Length") +
  theme_minimal()
```



There is a *clear increasing trend in median tooth length as the dose of vitamin C increases (from 0.5 mg to 2 mg).* The medians are visually distinct across the dose levels. As for the spread, it also changes with the dose. *The lowest dose (0.5 mg) has smaller variability, while the higher doses (1 mg and 2 mg) show wider ranges of tooth length.* These patterns suggest that dose level has a significant effect on tooth growth, with higher doses leading to greater growth in tooth length.

Q2: Additive two-way ANOVA model of len on dose and supp

```r
# Fit additive ANOVA model
additive_model <- aov(len ~ supp + dose, data = ToothGrowth)
```

```r
# Summary of the ANOVA
summary(additive_model)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## supp         1  205.4   205.4   14.02 0.000429 ***
## dose         2 2426.4  1213.2   82.81  < 2e-16 ***
## Residuals   56  820.4    14.7
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

P-value for supp is 0.000429, and the p-value for dose is smaller than 2e-16.

Q3: Interation model

```r
# Fit the interaction model
interaction_model <- aov(len ~ supp * dose, data = ToothGrowth)

# Summary of the interaction model
summary(interaction_model)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## supp         1  205.4   205.4  15.572 0.000231 ***
## dose         2 2426.4  1213.2  92.000  < 2e-16 ***
## supp:dose    2  108.3    54.2   4.107 0.021860 *
## Residuals   54  712.1    13.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Numerator degrees of freedom and P-value for interaction
anova(additive_model, interaction_model)
```

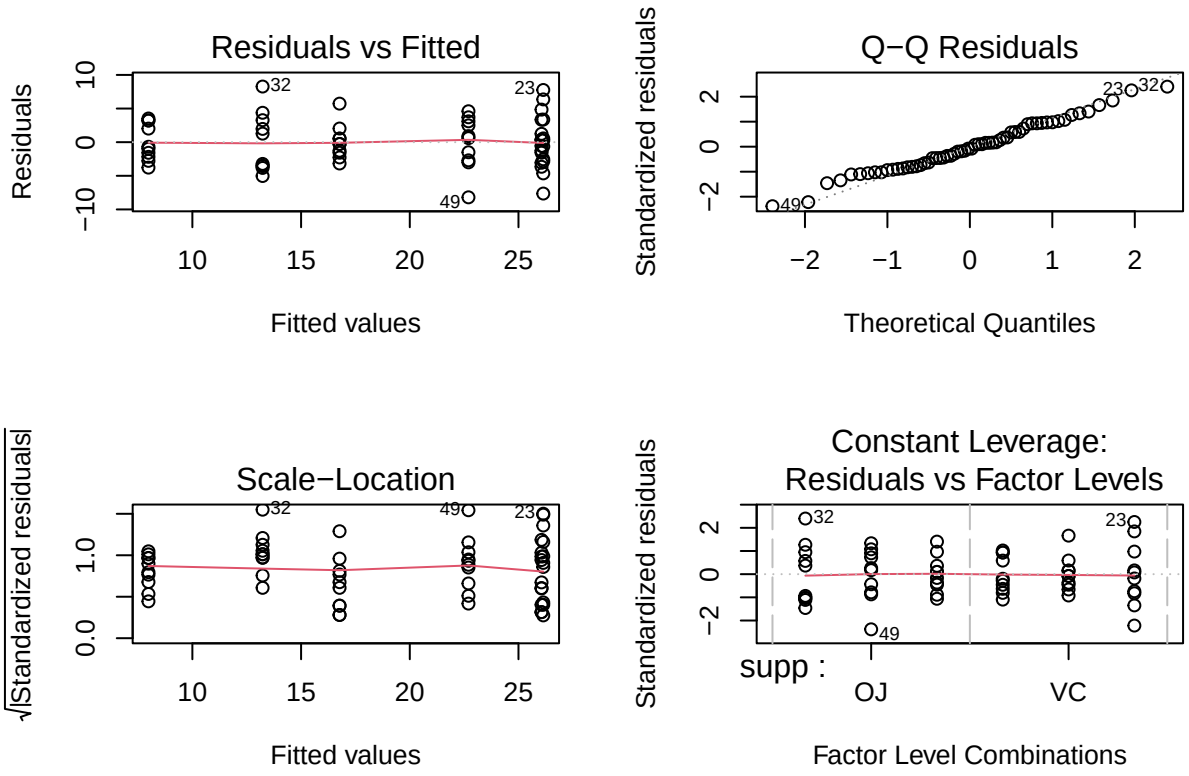
```
## Analysis of Variance Table
##
## Model 1: len ~ supp + dose
## Model 2: len ~ supp * dose
##   Res.Df    RSS Df Sum of Sq     F  Pr(>F)
## 1     56 820.43
## 2     54 712.11  2    108.32 4.107 0.02186 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As we can see from both the summary and the anova function output, the numerator degrees of freedom is 2, with a p-value of 0.02186.

Since this p-value is less than 0.05, the interaction is statistically significant at the 5% significance level. This means that the effect of dose on tooth length depends on the type of supplement and dose.

Q4: Diagnostics plots

```
par(mfrow = c(2, 2))
plot(interaction_model)
```



As for non-normality and non-constant variance, we will mainly look at the Scale-Location plt and the Normal Q-Q Plot.From the Scale-Location plt, the red line is relatively flat, and the residuals appear evenly spread across fitted values, suggesting that the assumption of constant variance is met. From the normal Q-Q Plot, most points lie close to the diagonal, except for deviations at the tails. This suggests slight non-normality in the residuals, which may or may not significantly impact the model.

(not that important)From the Residuals vs. Fitted plot, residuals appear fairly evenly scattered around zero, with no clear pattern, which suggests the model's linearity assumption holds.

```
# Cook's Distance
cooks_distance <- cooks.distance(interaction_model)

# find influential points
threshold <- 4 / nrow(ToothGrowth)
influential_points <- which(cooks_distance > threshold)

influential_points
```

```
## 22 23 32 49
## 22 23 32 49
```

We can tell the outliers from all four plots above(Residuals vs. Fitted etc), those outliers includes point 32, 23, 49. The Influence Plot *shows some influential points (#22, #23, #32, #49)* with noticeable influence (high leverage or residuals). These points should be carefully investigated as they could disproportionately affect the model's fit. The Cook's Distance values seem quite low, which suggests no points are excessively influential based on Cook's D threshold, but the flagged points (#22, #23, #32, #49) may still deserve further scrutiny.

Q5: Test Null Hypothesis: Dose Treated as Numerical

```
# Fit model treating dose as numerical
numerical_model <- aov(len ~ supp * as.numeric(as.character(dose)), data = ToothGrowth)

# Compare models using F-test
anova(interaction_model, numerical_model)
```

```
## Analysis of Variance Table
##
## Model 1: len ~ supp * dose
## Model 2: len ~ supp * as.numeric(as.character(dose))
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     54 712.11
## 2     56 933.63 -2   -221.53 8.3994 0.0006667 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value for the f-test is 0.0006667, which is much smaller than 0.05. Thus, we have sufficient evidence to reject the null hypothesis. Thus, dose should not be treated as numerical (len supp*dose)), and it is more appropriate to treat dose as a categorical variable in the analysis. The hypothesis in model 3 is correct.