"Inferential Data Analysis Project"

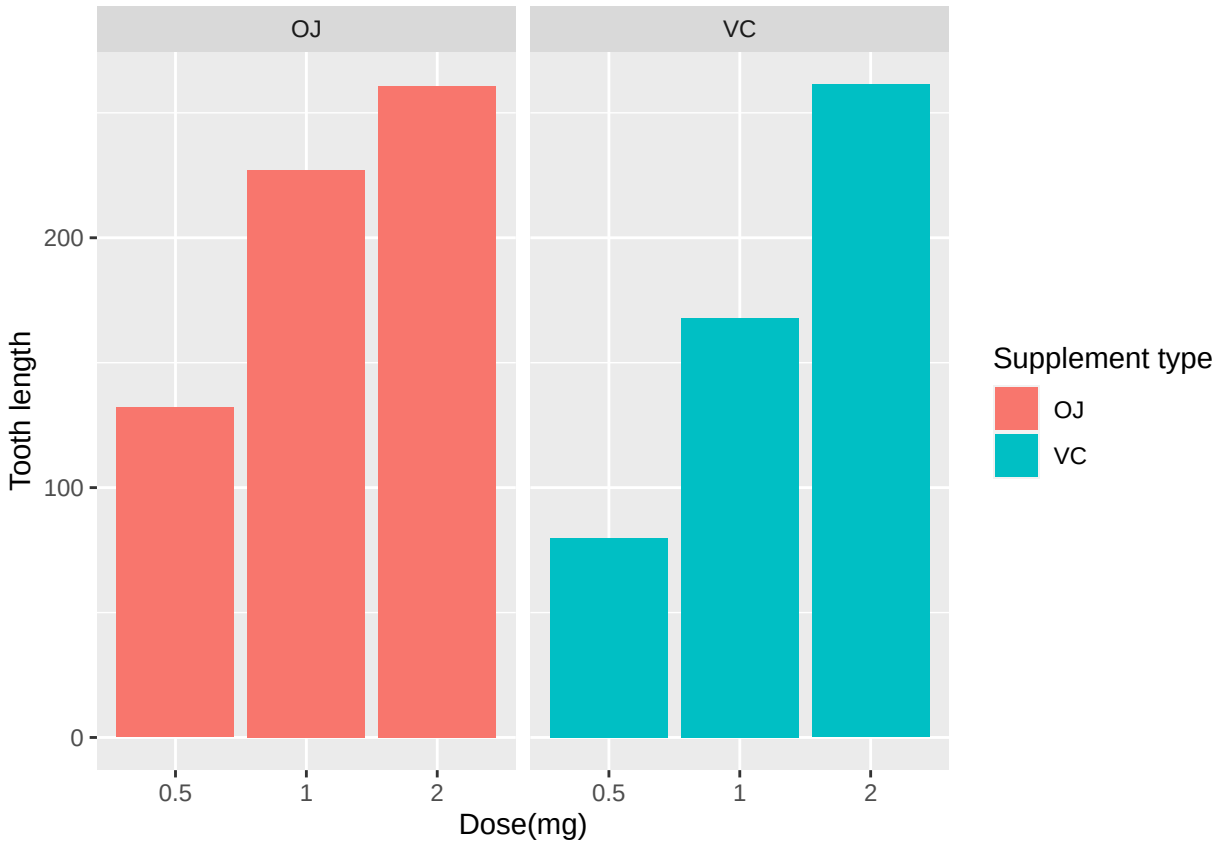
Tiffany M. Kollah

5.18.2020

```
library(datasets)
library(ggplot2)
```

##Plot 1

```
ggplot(data=ToothGrowth, aes(x=as.factor(dose), y=len, fill=supp)) +
  geom_bar(stat="identity",) +
  facet_grid(. ~ supp) +
  xlab("Dose(mg)") +
  ylab("Tooth length") +
  guides(fill=guide_legend(title="Supplement type"))
```



Fit linear model

```
fit <- lm(len ~ dose + supp, data=ToothGrowth)
```

##Get the summary of the Data Set

```
summary(fit)
```

```
##
## Call:
## lm(formula = len ~ dose + supp, data = ToothGrowth)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.600 -3.700  0.373  2.116  8.800
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.2725     1.2824   7.231 1.31e-09 ***
## dose          9.7636     0.8768  11.135 6.31e-16 ***
## suppVC       -3.7000     1.0936  -3.383   0.0013 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.236 on 57 degrees of freedom
## Multiple R-squared:  0.7038, Adjusted R-squared:  0.6934
## F-statistic: 67.72 on 2 and 57 DF,  p-value: 8.716e-16
```

##As you see, the model explains 70% of the variance in the data. The intercept is r fit$coefficients[[1]], $meaningthatwithnosu$ units. The coefficient of dose is r fit$coefficients[[2]]. It can be interpreted as increasing the delievered dose 1 mg, $allelseequal(i.e.n$ units. The last coefficient is for the supplement type. Since the supplement type is a categorical variable, dummy variables are used. The computed coefficient is for suppVC and the value is r fit$coefficients[[3]] $meaningthatdeliveringagivendoseasascorbicacid, withoutchangingthedose, wouldresultinrabs(fitcoeffici$ units of decrease in the tooth length. Since there are only two categories, we can also conclude that on average, delivering the dosage as orange juice would increase the tooth length by r abs(fit$coefficients[[3]]) units.95% confidence intervals for two variables and the intercept are as follows.

```
confint(fit)
```

```
##                 2.5 %    97.5 %
## (Intercept)  6.704608 11.840392
## dose         8.007741 11.519402
## suppVC      -5.889905 -1.510095
```

##Conclusion:The confidence intervals mean that if we collect a different set of data and estimate parameters of the linear model many times, 95% of the time, the coefficient estimations will be in these ranges. For each coefficient (i.e. intercept, dose and suppVC), the null hypothesis is that the coefficients are zero, meaning that no tooth length variation is explained by that variable. All p-values are less than 0.05, rejecting the null hypothesis and suggesting that each variable explains a significant portion of variability in tooth length, assuming the significance level is 5%.