Welcome! Introduction to Statistics using R

"Researchers should use statistics in the same way that a drunk uses lampposts—for support rather than illumination"

-Andrew Lang (modified quote by unknown)[1]

[1] Original quote: "Politicians use statistics in the same way that a drunk uses lampposts, for support rather than illumination" -Andrew Lang

Outline

Links for getting started

Download R and Rstudio, respectively:
https://cran.r-project.org/index.html
https://rstudio.com/products/rstudio/download/

Quick guides for R:
https://rstudio.com/resources/cheatsheets/

Book - Hadley Wickham's R for data science:
https://r4ds.had.co.nz/

Swirl - Interactive learning for R:
https://swirlstats.com/

Input data

R Code

```
#create first column
fruit <- c("apple", "apple", "apple", "orange", "orange", "orange")

#create second column (1 = youngest, 5 = oldest)
age <- as.numeric(c(2,2,2,3,4,5))

#create dataframe
fruits <- data.frame(fruit,age)

#save dataframe to computer hard drive
write.table(fruits, file="fruitsSTATSU.csv", sep= ",",
row.names=FALSE,
col.names=TRUE)
```

Import file: *dataset <- read.csv("path/path/file.csv")*

Simulate data

R Code

```
#set.seed for reproducibility
set.seed(300)

#randomly sample 500 obs. with a mean of 0, SD of 1
Samp1 <- rnorm(500, mean=0, sd=1)
#randomly sample 500 obs. with a mean of 1, SD of 2
Samp2 <- rnorm(500, mean=1, sd=2)
#randomly sample 500 obs. with a mean of 2, SD of 3
Samp3 <- rnorm(500, mean=2, sd=3)

#combine samples into a data frame
data <- data.frame(Samp1, Samp2, Samp3)
```

Descriptive statistics

R Code

psych::describe(data)

Descriptive statistics

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
Samp1	1	500	0.07	1.01	0.07	0.06	1.04	-3.14	3.50	6.65	0.06	0.01	0.05
Samp2	2	500	1.01	2.09	0.99	1.02	2.04	-5.34	7.02	12.36	-0.06	-0.10	0.09
Samp3	3	500	2.36	2.95	2.50	2.37	3.06	-6.85	10.63	17.48	-0.05	-0.28	0.13

Interquartile range

R Code

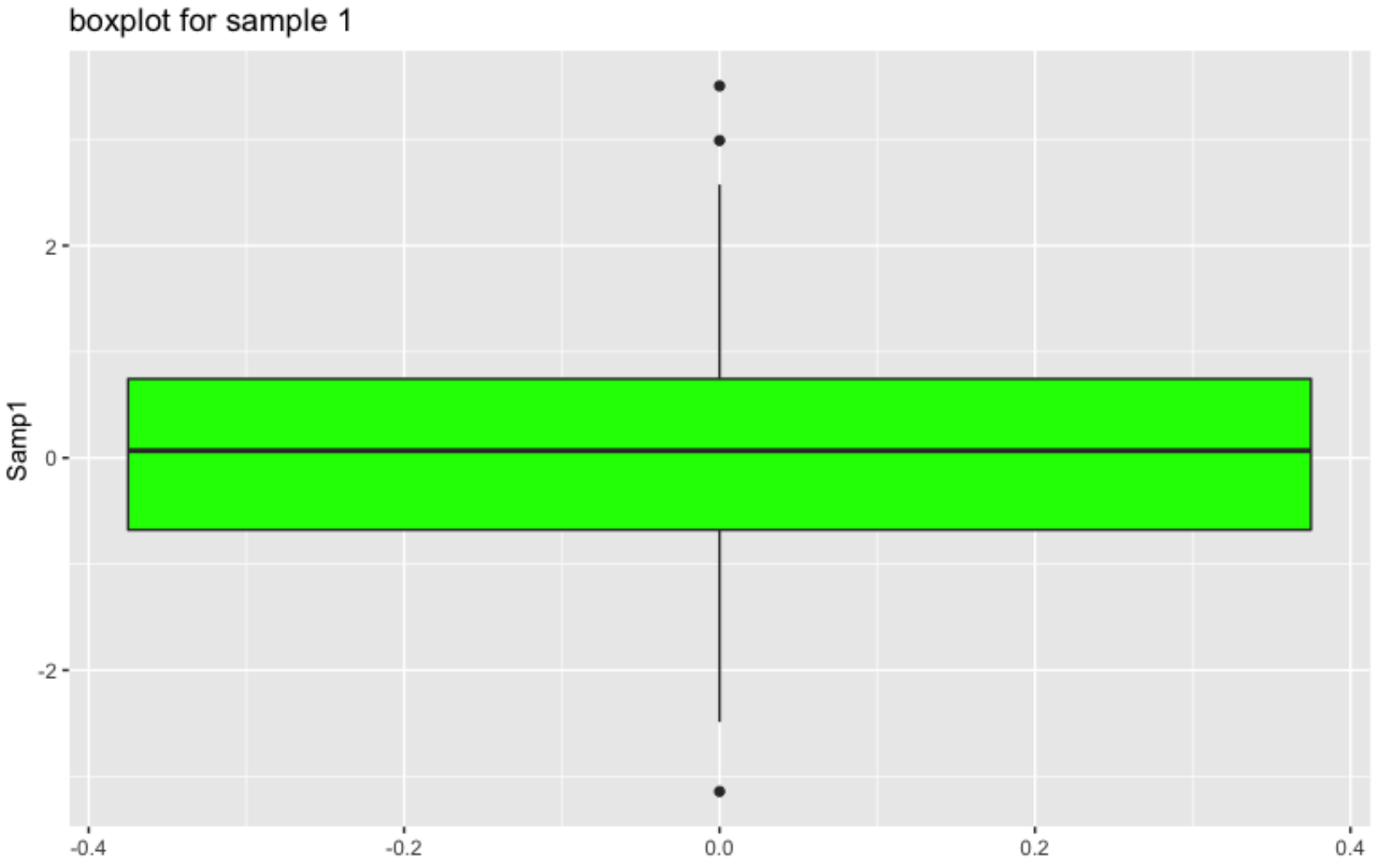
```
#Find interquartile range:
IQR(Samp1)

#Find quartiles of a vector:
quantile(Samp1)

##Find quartiles of one variable column within a dataframe:
quantile(data$Samp1)
```

Boxplot code

R Code

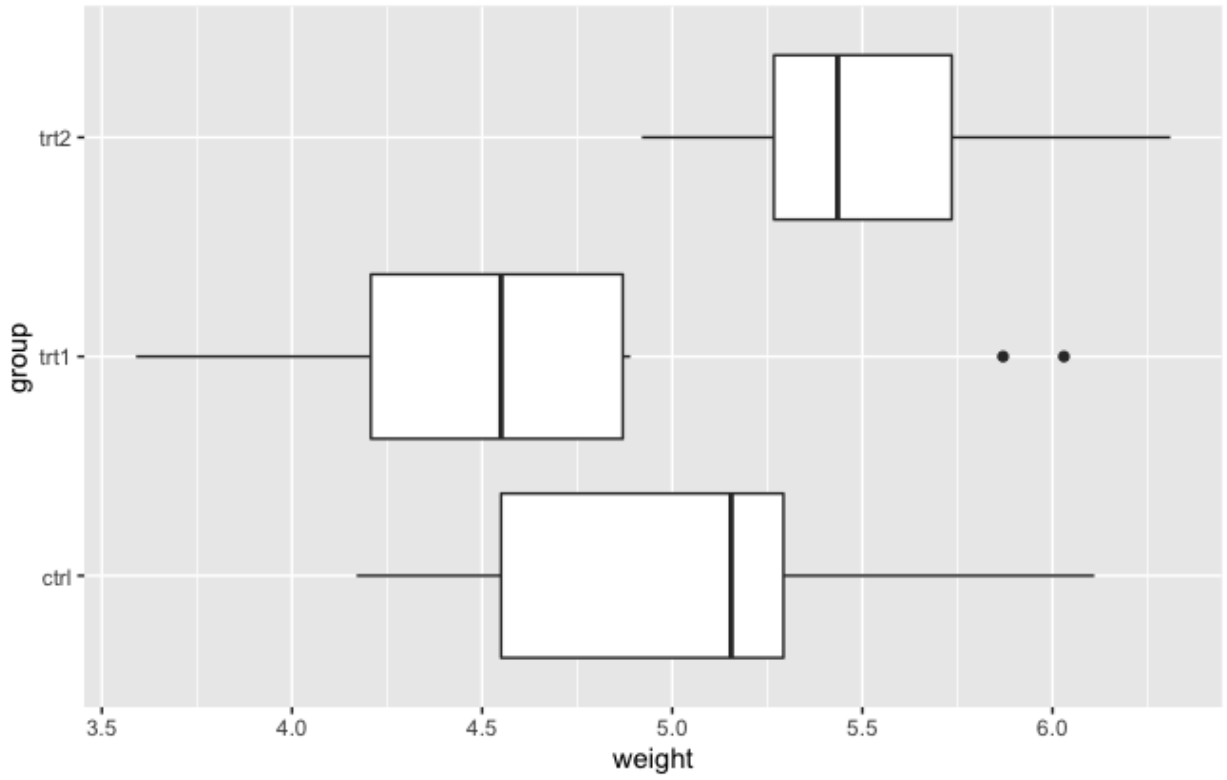
```
#boxplot ggplot(data, aes(y=Samp1)) +
geom_boxplot(fill="green") +
labs(title="boxplot for sample 1")
```

Boxplot and quartiles ouptut



0%	25%	50%	75%	100%
-3.1425972	-0.6781704	0.0678328	0.7443345	3.5041409

Boxplot for PlantGrowth

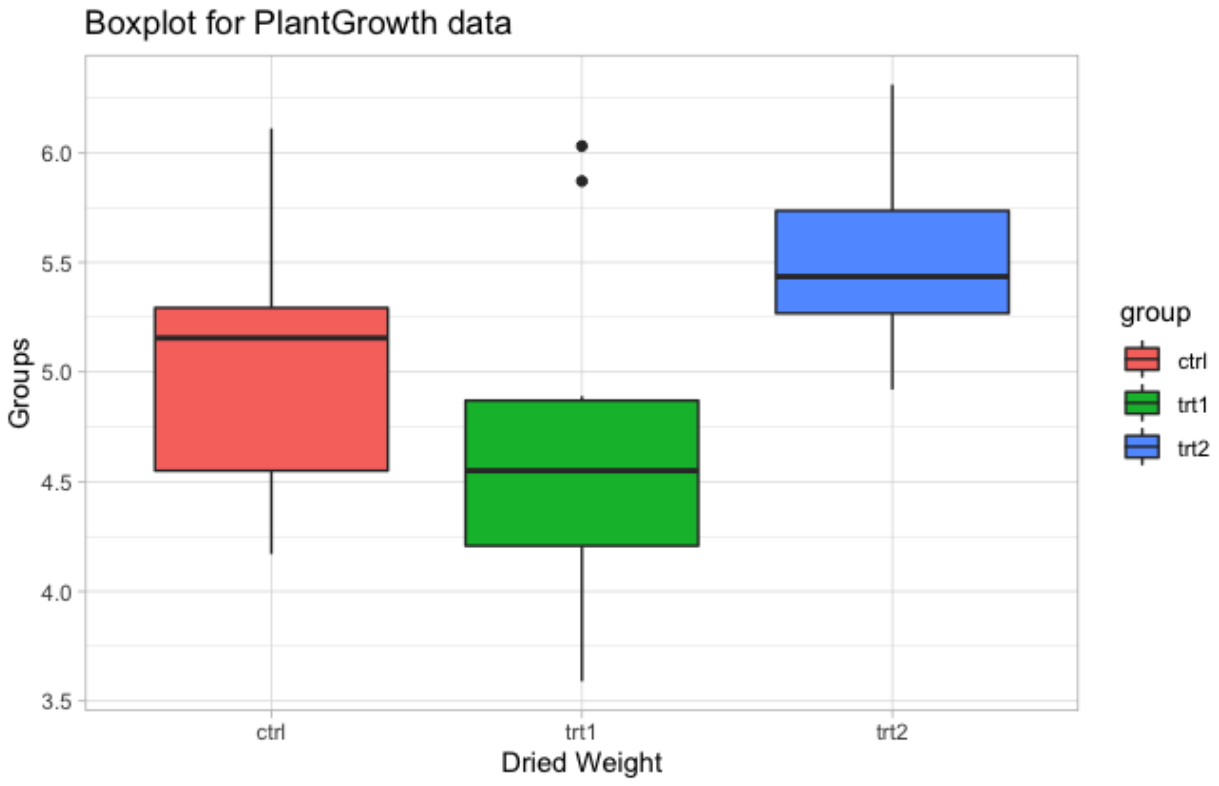


R Code

```
ggplot(PlantGrowth, aes(x=weight, y=group)) +
geom_boxplot()
```

Boxplot for PlantGrowth

modify the boxplot (code on following page):



Boxplot modifications

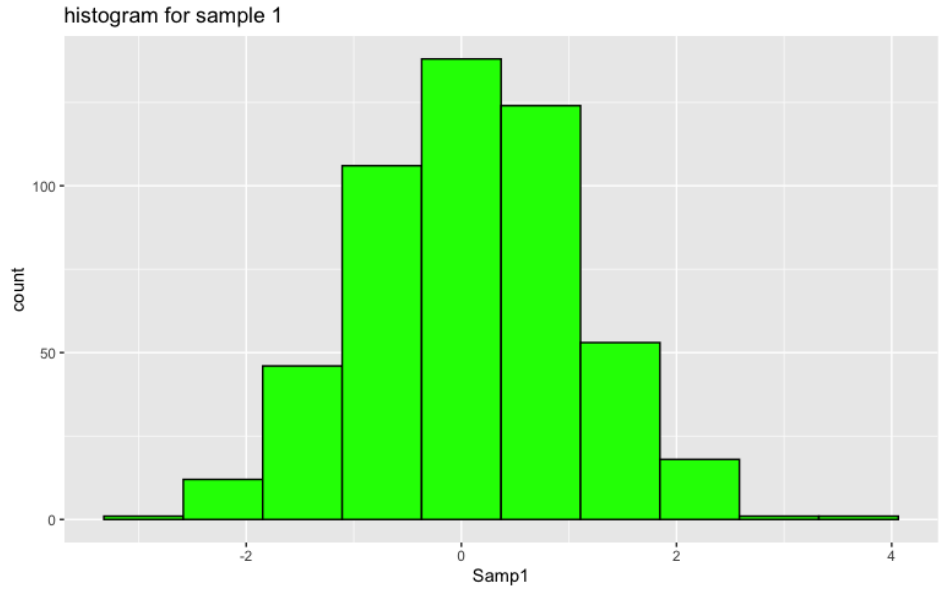
R Code

```
ggplot(PlantGrowth, aes(x=weight, y=group, fill=group)) +
geom_boxplot()+
labs(x= "Groups", y= "Dried Weight", title="Boxplot for
PlantGrowth data") +
coord_flip()+
theme_light()
```

Histogram

R Code

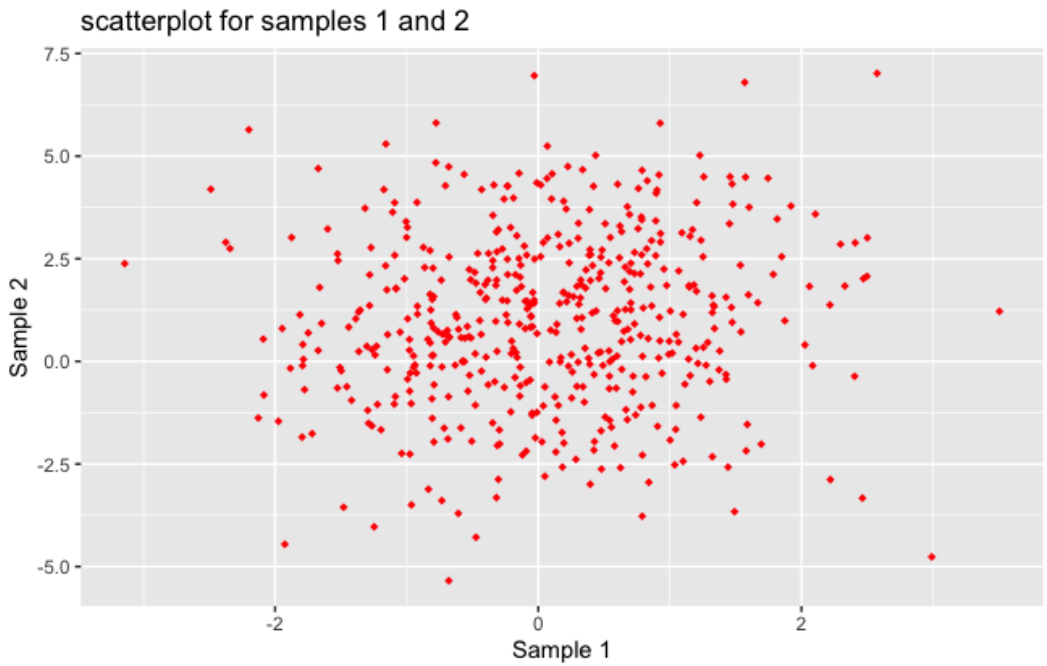
```
#histogram ggplot(data, aes(x=Samp1)) +
geom_histogram(bins=10, color="black", fill="green") +
labs(title="histogram for sample 1")
```



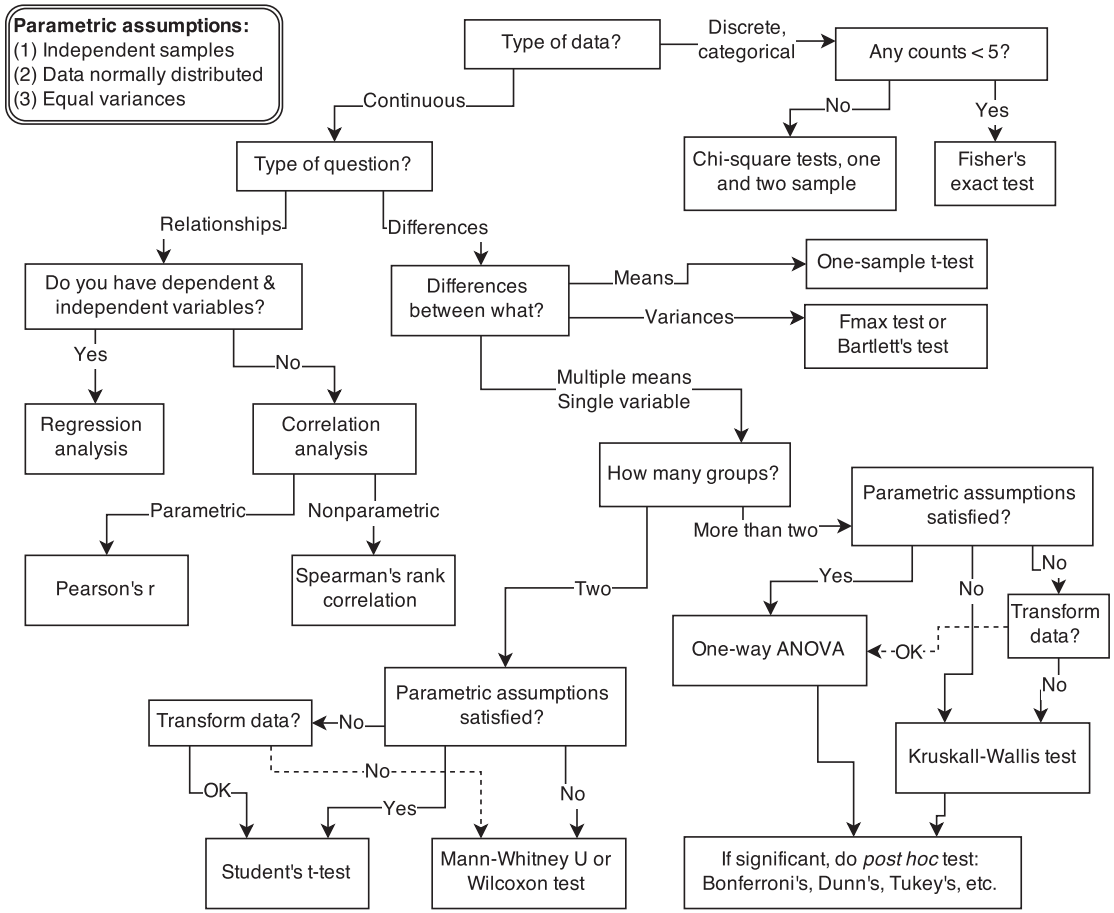
Scatterplot

R Code

```
#scatterplot
ggplot(data, aes(x=Samp1, y=Samp2)) +
geom_point(shape=18, color="red") +
labs(x= "Sample 1", y= "Sample 2" ,title="scatterplot for samples
1 and 2")
```



Statistics flow chart



Statistical Rethinking: A Bayesian Course with Examples in R and Stan by Richard McElreath

Correlation

R Code

cor(data$var1, data$var2)
cor.test(data$var1, data$var2)

```
> cor(data$Samp1, data$Samp2)
[1] 0.08818687
> cor.test(data$Samp1, data$Samp2)

        Pearson's product-moment correlation

data:  data$Samp1 and data$Samp2
t = 1.9757, df = 498, p-value = 0.04874
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.0005001525 0.1745278088
sample estimates:
       cor
0.08818687
```

t-tests

R Code

```
#One sample t-test
t.test(data$Samp1, mu=0)

#Two-sample independent t-test
t.test(data$Samp1, data$Samp2, var.equal=TRUE,
paired=FALSE)

#Two-sample dependent t-test
t.test(data$Samp1, data$Samp2, var.equal=FALSE,
paired=TRUE)
```

χ^2 test

Goodness of fit

```
> group1 <- 300
> group2 <- 350
> group3 <- 100
> data <- c(group1, group2, group3)
>   chisq.test(data)

        Chi-squared test for given probabilities

data:  data
X-squared = 140, df = 2, p-value < 2.2e-16
```

ANOVA

One-Way ANOVA

```
> var1 <- rnorm(12, mean=2, sd=1)
> var2 <- c("TX1", "TX1", "TX1", "TX1",
+           "C", "C", "C", "C", "C",
+           "TX2","TX2", "TX1")
> ANOVAdat <- data.frame(var1, var2)
> model <- aov(ANOVAdat$var1 ~ ANOVAdat$var2, data=ANOVAdat)
> model
Call:
   aov(formula = ANOVAdat$var1 ~ ANOVAdat$var2, data = ANOVAdat)

Terms:
                ANOVAdat$var2 Residuals
Sum of Squares       6.910097 11.089850
Deg. of Freedom             2         9

Residual standard error: 1.110048
Estimated effects may be unbalanced
```

ANOVA

Two-Way ANOVA

```
> var1 <- rnorm(12, mean=2, sd=1);
> var2 <- c("ONE", "ONE", "ONE", "ONE", "TWO", "TWO",
+           "TWO", "TWO", "TWO", "THREE","THREE", "ONE")
> var3 <- c("THREE", "THREE", "THREE", "THREE", "FOUR",
+           "FOUR", "FOUR", "FOUR", "FOUR", "FIVE", "FIVE", "FIVE")
> ANOVAdat <- data.frame(var1, var2, var3)
> model2 <- aov(ANOVAdat$var1 ~ ANOVAdat$var2 + ANOVAdat$var3, data=ANOVAdat)
> model2
Call:
   aov(formula = ANOVAdat$var1 ~ ANOVAdat$var2 + ANOVAdat$var3,
    data = ANOVAdat)

Terms:
                ANOVAdat$var2 ANOVAdat$var3 Residuals
Sum of Squares       0.568126      2.638690  7.286094
Deg. of Freedom             2             1         8

Residual standard error: 0.9543384
1 out of 5 effects not estimable
Estimated effects may be unbalanced
```

Regression

```
> IV <- rnorm(300, mean=1, sd=3)
> DV <- rnorm(300, mean=0, sd=5)
> Rdata <- data.frame(IV, DV)
> model3 <- lm(data$DV ~ data$IV, data=Rdata)
> model3;

Call:
lm(formula = data$DV ~ data$IV, data = Rdata)

Coefficients:
(Intercept)       data$IV
  -0.388191      0.003915
```

Regression

summary() function

```
> summary(model3)

Call:
lm(formula = data$DV ~ data$IV, data = Rdata)

Residuals:
     Min       1Q   Median       3Q      Max
-12.6540  -3.2726   0.2452   3.2431  16.7492

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.388191   0.297573  -1.305    0.193
data$IV      0.003915   0.094223   0.042    0.967

Residual standard error: 5.005 on 298 degrees of freedom
Multiple R-squared:  5.793e-06, Adjusted R-squared:  -0.00335
F-statistic: 0.001726 on 1 and 298 DF,  p-value: 0.9669
```

Alternative modeling code

R Code

```
IV <- rnorm(300, mean=1, sd=3)
DV <- rnorm(300, mean=0, sd=5)
data <- data.frame(IV, DV)
model <- lm(data$DV ~ data$IV, data=data)
summary(model)

# Type I
anova(model)
# Type II
car::Anova(model)
#Type III
options(contrasts=c("contr.sum","contr.poly"))
car::Anova(model, type=3)
```

Thank you!

Questions?