An Introduction to *R*

Part 2: Comparing Groups

Kjell Johnson

Overview

- Overview of one-way analysis of variance (ANOVA)
- Data shaping: How to put the data into the form we need prior to analysis
- Visualizing the data
- Performing one-way ANOVA
  - Checking assumptions
- Pairwise treatment comparisons (what we *really* want to know)

Comparing Multiple Groups

- Common situations:
  - multiple compounds
  - multiple doses of same compound
- Control groups:
  - negative control (e.g. saline)
  - untreated control
  - positive control: typically a compound with a known effect, often used for assay validation
- Idea: use data from all groups to make comparisons

Analysis of Variance (ANOVA)

- ANOVA is used to compare multiple groups

- Assumptions:
  - Data are independent
    - not repeated measures or measure replicates
  - Residuals are normally distributed
  - Group variances are similar

ANOVA Hypotheses and Test

- Null hypothesis: all group means are equal

$$H_0\text{:}\ \ \mu_1 = \mu_2 = \ldots = \mu_r$$

- Alternative hypothesis: at least one pair of group means are different

$$H_A\text{:}\ \ \mu_k \neq \mu_l$$

- How do we know? **ANOVA F-test**
  - The F-statistic is a *variance ratio*
  - Thus, "Analysis of *Variance*"

ANOVA F Test Statistic

Variability among data

=

Variability between groups

+

Variability within groups

$$\text{F - statistic} = \frac{\text{Variability Between Groups}}{\text{Variability Within Groups}}$$

Example: Colon Cancer Chemoprevention

50 animals were randomly divided into 5 groups. Each group received a different treatment. Are the tumor diameters different among treatments?

Animal	Control	Drug_A	Drug_B	Drug_C	Drug_D
1	2.27	1.73	0.97	1.29	0.50
2	1.38	1.19	1.08	1.13	0.70
3	1.91	1.39	0.77	1.12	1.55
4	2.21	1.08	1.29	1.08	0.98
5	2.63	1.14	1.08	1.71	0.65
6	2.73	1.22	1.18	2.49	0.70
7	2.08	1.62	0.87	2.04	1.13
8	2.92	1.03	0.89	2.59	0.60
9	2.78	2.64	1.70	2.63	0.57
10	1.87	1.49	2.30	1.89	0.78

The Statistical Model

$$Y_{i,j} = \mu_i + \epsilon_{i,j}$$

- $Y_{i,j}$ is the observed response value for the j^{th} subject on the i^{th} treatment
  - i = 1,…, # treatments (5 for the tumor diameter example)
  - j = 1,…,# of subjects (10 for the tumor diameter example)
- μ_i is the effect of the i^{th} treatment
- $\varepsilon_{i,j}$ is the random effect for the j^{th} subject on the i^{th} treatment that is not explained by the i^{th} treatment effect.
  - The errors are independent and follow a normal distribution with constant variance.

Tumor Diameter Data (mm)

	Trt 1	Trt 2	Trt 3	Trt 4	Trt 5
Animal	**Control**	**Drug_A**	**Drug_B**	**Drug_C**	**Drug_D**
1	2.27	1.73	0.97	1.29	0.50
2	1.38	1.19	1.08	1.13	0.70
3	1.91	1.39	0.77	1.12	1.55
4	2.21	1.08	1.29	1.08	0.98
5	2.63	1.14	1.08	1.71	0.65
6	2.73	1.22	1.18	2.49	0.70
7	2.08	1.62	0.87	2.04	1.13
8	2.92	1.03	0.89	2.59	0.60
9	2.78	2.64	1.70	2.63	0.57
10	1.87	1.49	2.30	1.89	0.78

$Y_{1,1}$

$Y_{1,2}$

$Y_{1,10}$

$Y_{5,10}$

Bring the Data into R

Set the working directory:

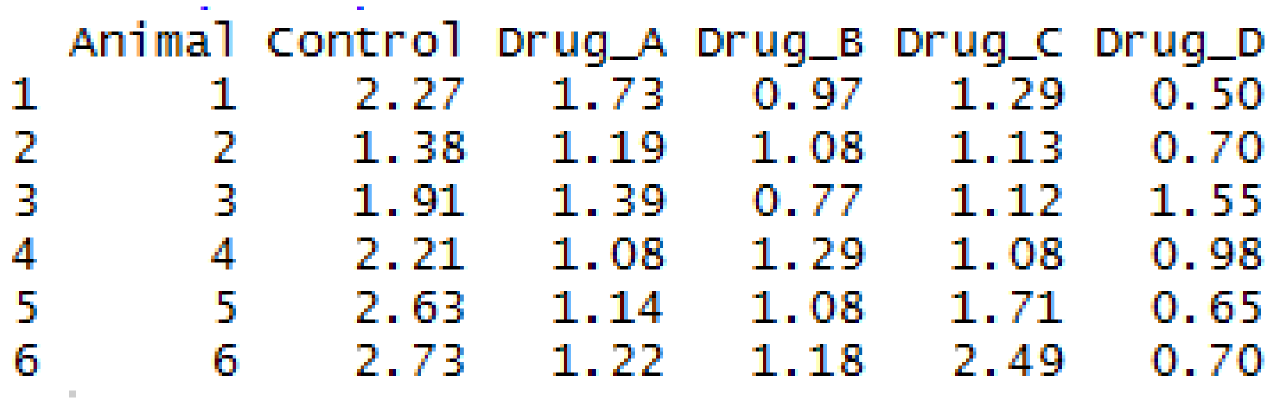
```
myLocation <- "c:/Documents and Settings/johns94/Desktop/Part2"
setwd(myLocation)
```

Get data:

```
tumor <- read.csv("tumor.csv",header=TRUE)
```

Look at the top of the file:

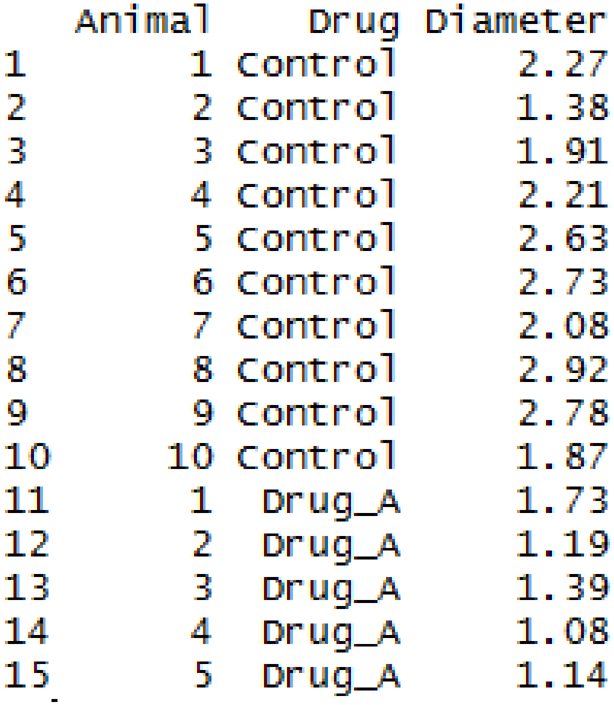
```
head(tumor)
```



```
  Animal Control Drug_A Drug_B Drug_C Drug_D
1      1    2.27   1.73   0.97   1.29   0.50
2      2    1.38   1.19   1.08   1.13   0.70
3      3    1.91   1.39   0.77   1.12   1.55
4      4    2.21   1.08   1.29   1.08   0.98
5      5    2.63   1.14   1.08   1.71   0.65
6      6    2.73   1.22   1.18   2.49   0.70
```

“Wide” Versus “Narrow” Files

- We call this a “wide” file, since each drug is in a separate column. We need the data to be in “narrow” form where the treatment information is in a column and the response is in another column:



```
   Animal    Drug Diameter
1       1 Control     2.27
2       2 Control     1.38
3       3 Control     1.91
4       4 Control     2.21
5       5 Control     2.63
6       6 Control     2.73
7       7 Control     2.08
8       8 Control     2.92
9       9 Control     2.78
10     10 Control     1.87
11      1  Drug_A     1.73
12      2  Drug_A     1.19
13      3  Drug_A     1.39
14      4  Drug_A     1.08
15      5  Drug_A     1.14
```

Data Shaping

- We could manually cut-and-paste to get this form, but that's tedious and prone to mistakes.
- Good news! We can transform the shape of the data directly in R.

Install and load the reshape package:

```
install.packages("reshape", dependencies=TRUE)
library(reshape)
```

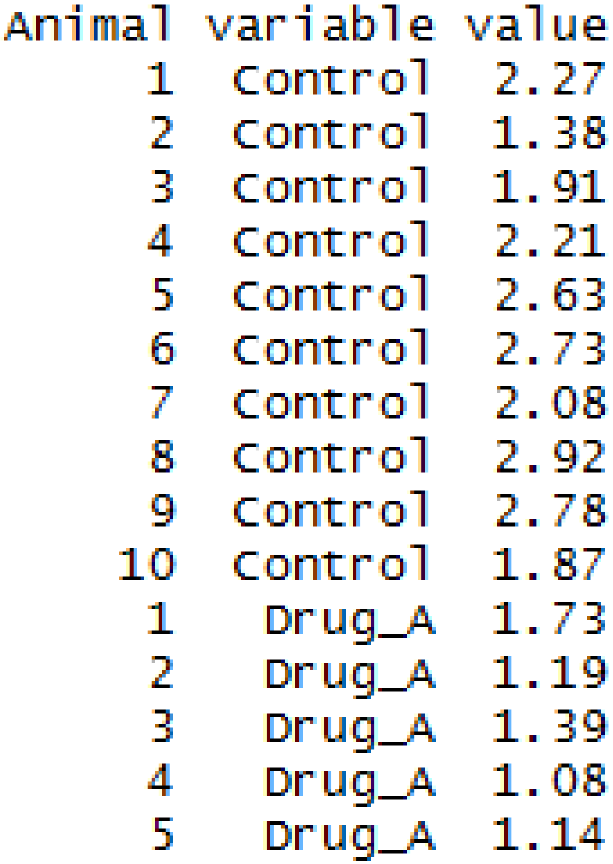
Transform data to narrow form using the melt function in reshape:

```
tumorNarrow = melt(tumor,id="Animal")
```

The id option tells the function the variables that should be kept in the "stacking" process.

Results of “melt”

First 15 rows of reshaped data:



```
Animal variable value
     1  Control  2.27
     2  Control  1.38
     3  Control  1.91
     4  Control  2.21
     5  Control  2.63
     6  Control  2.73
     7  Control  2.08
     8  Control  2.92
     9  Control  2.78
    10  Control  1.87
     1   Drug_A  1.73
     2   Drug_A  1.19
     3   Drug_A  1.39
     4   Drug_A  1.08
     5   Drug_A  1.14
```

Rename columns:

```
colnames(tumorNarrow)[2:3] <- c("Drug","Diameter")
```

Visualize Data

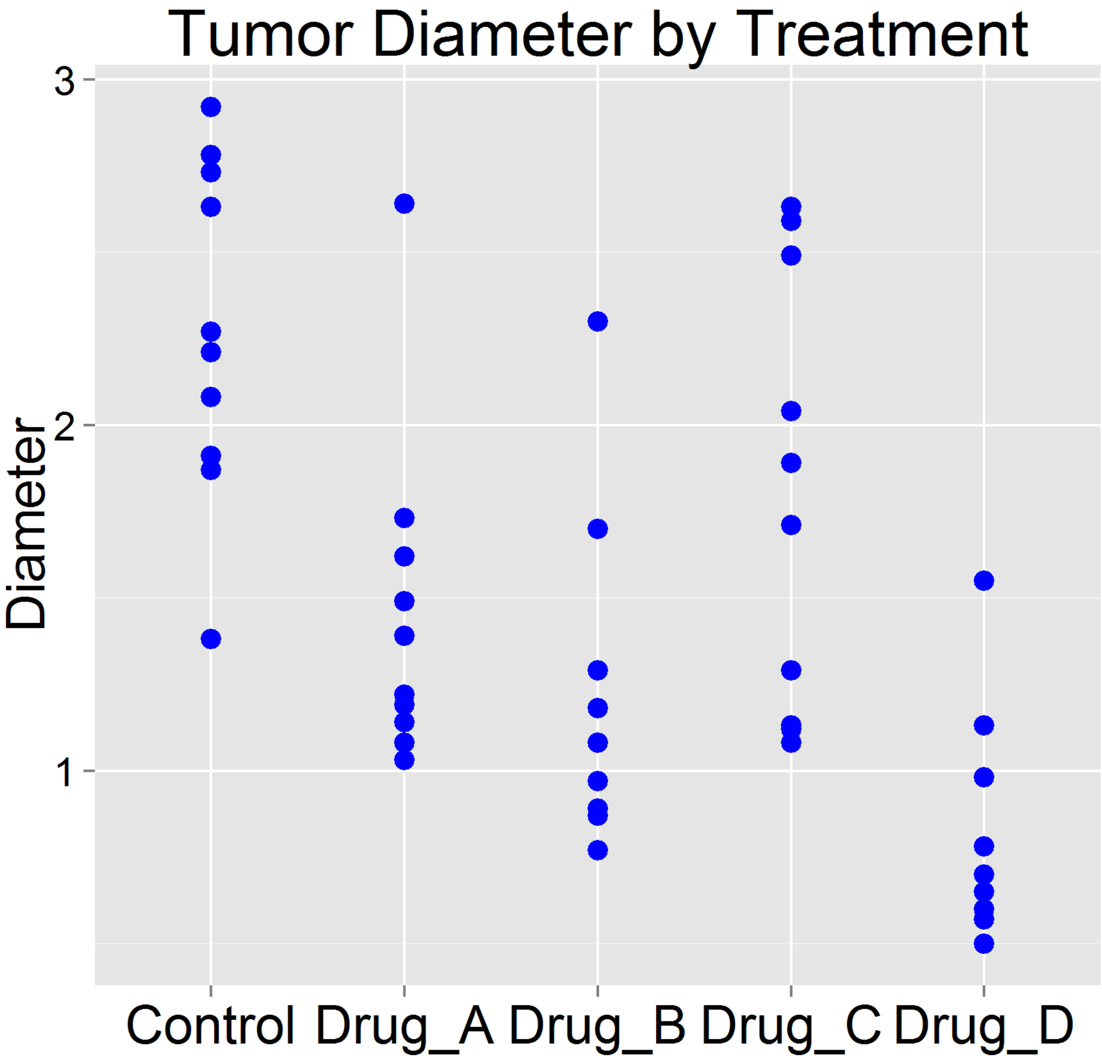
Create plot with Drug treatment on the x-axis and tumor diameter on the y-axis:

```
library(ggplot2)
ggplot(tumorNarrow, aes(x=Drug,y=Diameter)) +
  geom_point(color = "blue",
             shape = 20,
             size = 5) +
  ggtitle("Tumor Diameter by Treatment") +
  ylab("Diameter") +
  xlab("") +
  theme(axis.text.x = element_text(size=20,color="black"),
        axis.text.y = element_text(size=15,color="black"),
        axis.title.y = element_text(size=20),
        title = element_text(size=20))
```

Save the graph:

```
ggsave(file = "tumorFigure1.png")
```

Initial Figure



Modify Plot to Add Means and SD's

- We often want to see means and SD's on the figure. To do that we must first compute these values.

Install and load the doBy package:

```
install.packages("doBy", dependencies=TRUE)
library(doBy)
```

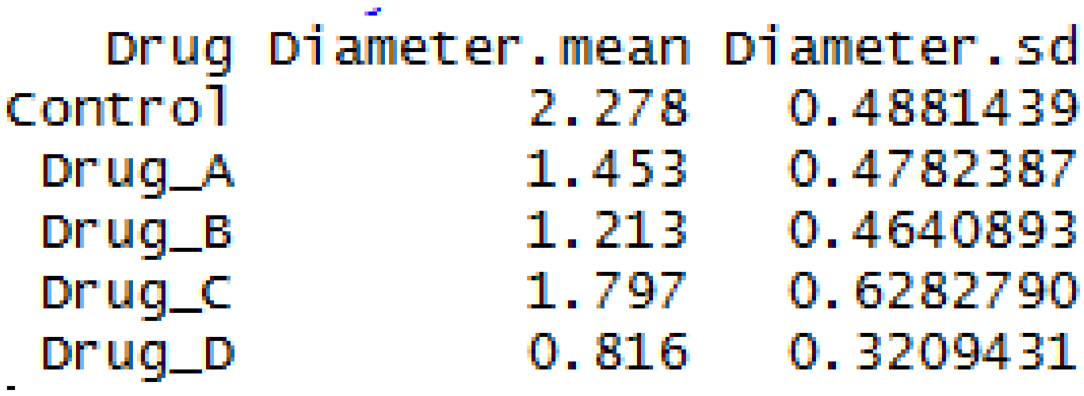
Compute means and standard deviations of Diameter for each Drug:

```
tumorSummary <- summaryBy(Diameter ~ Drug,
                          data = tumorNarrow,
                          FUN = c(mean,sd))
```

The first line is a "formula" (more later), and the third line identifies which summary functions we want to use.

tumorSummary

- A new data frame with contents:



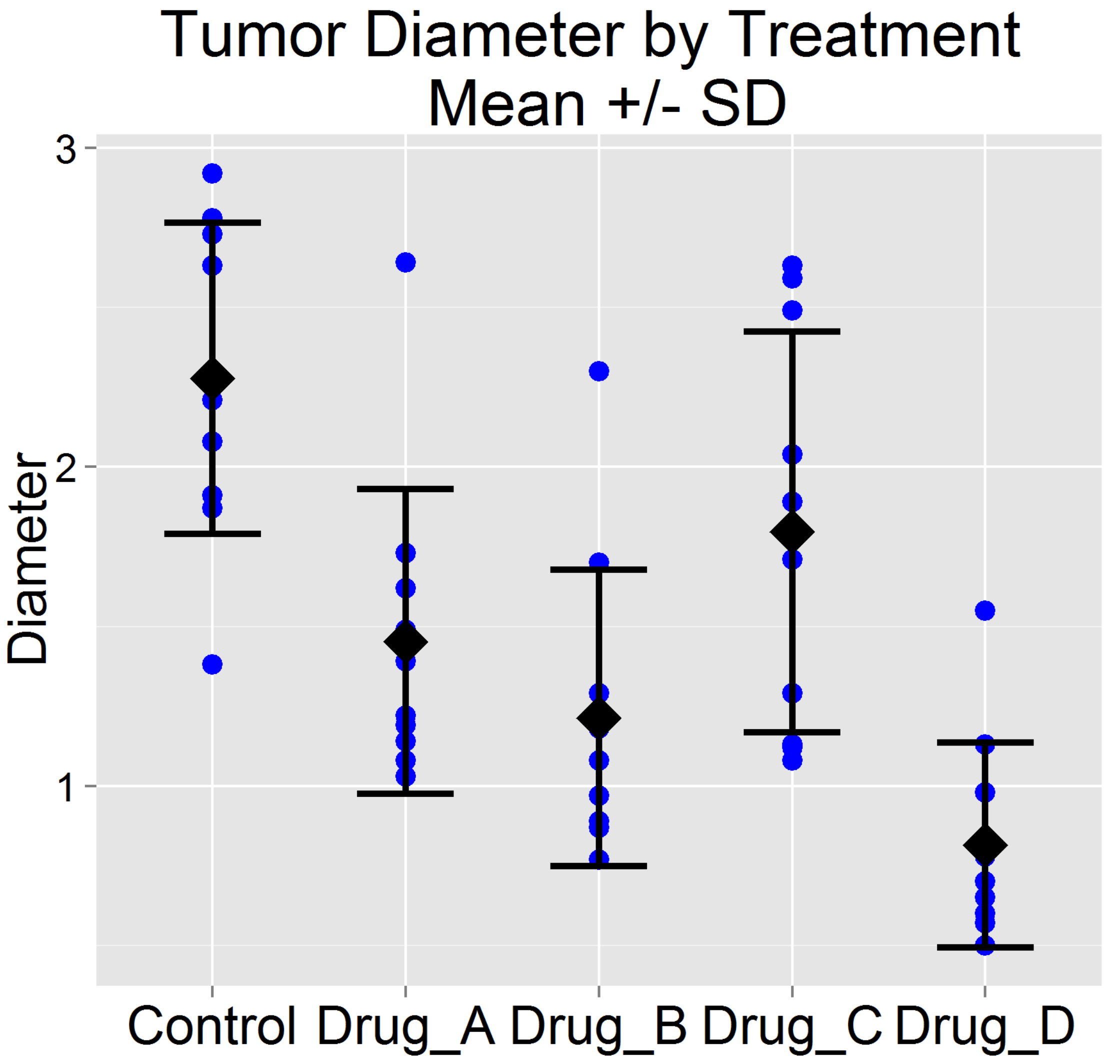
```
     Drug Diameter.mean Diameter.sd
Control           2.278   0.4881439
 Drug_A           1.453   0.4782387
 Drug_B           1.213   0.4640893
 Drug_C           1.797   0.6282790
 Drug_D           0.816   0.3209431
```

We now want to add this content to the figure.

New Figure

```
ggplot(data = tumorNarrow, aes(x=Drug,y=Diameter)) +
  geom_point(colour = "blue", shape = 20, size = 5) +
  ggtitle("Tumor Diameter by Treatment \n Mean +/- SD") +
  ylab("Diameter") + xlab("") +
  theme(axis.text.x = element_text(size=20,color="black"),
        axis.text.y = element_text(size=15,color="black"),
        axis.title.y = element_text(size=20),
        title = element_text(size=20))+
  geom_point(data=tumorSummary, aes(x=Drug, y=Diameter.mean),
             color = "black", size = 8, shape = 18) +
  geom_errorbar(data = tumorSummary,
                aes(x = Drug, y = Diameter.mean,
                    ymin = Diameter.mean - Diameter.sd,
                    ymax = Diameter.mean + Diameter.sd),
                color = "black",
                width = 0.5,
                size = 1.2)

ggsave(file = "tumorFigure2.png")
```

Updated Figure



The Formula Interface

- To estimate the parameters of a statistical model, R uses the “formula” interface.
- This interface places the “response” variable on the left and the “predictors” on the right:

$$y \sim x_1 + x_2 \ldots + x_p$$

- For this data, the response is Diameter and the predictor is Drug:

$$Diameter \sim Drug\ - 1$$

Technical Detail:

-1 means that there is no “intercept” in our model

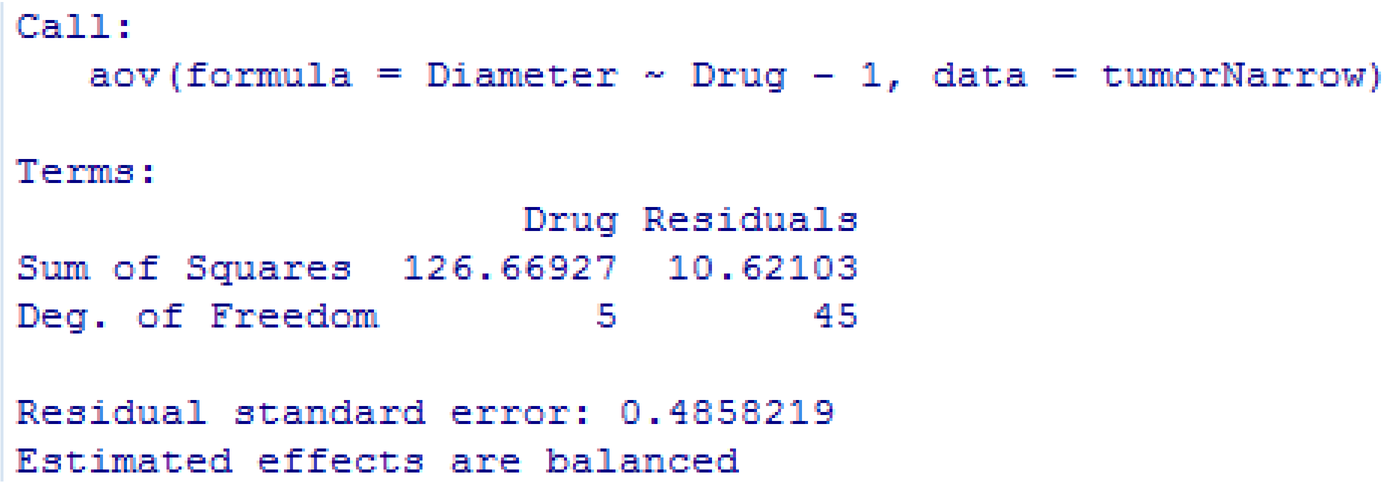
How to Perform ANOVA in R?

- The “aov” function performs analysis of variance and is a standard function in R (no additional package needed).

Perform ANOVA:

```
tumorANOVA = aov(Diameter ~ Drug - 1, data=tumorNarrow)
```

tumorANOVA is an object that contains all of the information about the model. Typing tumorANOVA at the prompt gives the following output:

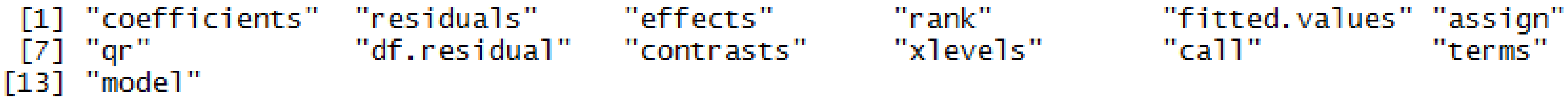


```
Call:
   aov(formula = Diameter ~ Drug - 1, data = tumorNarrow)

Terms:
                     Drug Residuals
Sum of Squares  126.66927  10.62103
Deg. of Freedom         5        45

Residual standard error: 0.4858219
Estimated effects are balanced
```

Objects in tumorANOVA

Get the names of the objects in tumorANOVA

```
names(tumorANOVA)
```



```
 [1] "coefficients"  "residuals"     "effects"       "rank"          "fitted.values" "assign"
 [7] "qr"            "df.residual"   "contrasts"     "xlevels"       "call"          "terms"
[13] "model"
```

There are other built-in R functions that call these objects and help us understand the residuals and the significance of the F-test. The "plot" function displays residual diagnostics:

```
plot(tumorANOVA)
```

The "summary" function provides the ANOVA F-test and corresponding signficance:

```
summary(tumorANOVA)
```

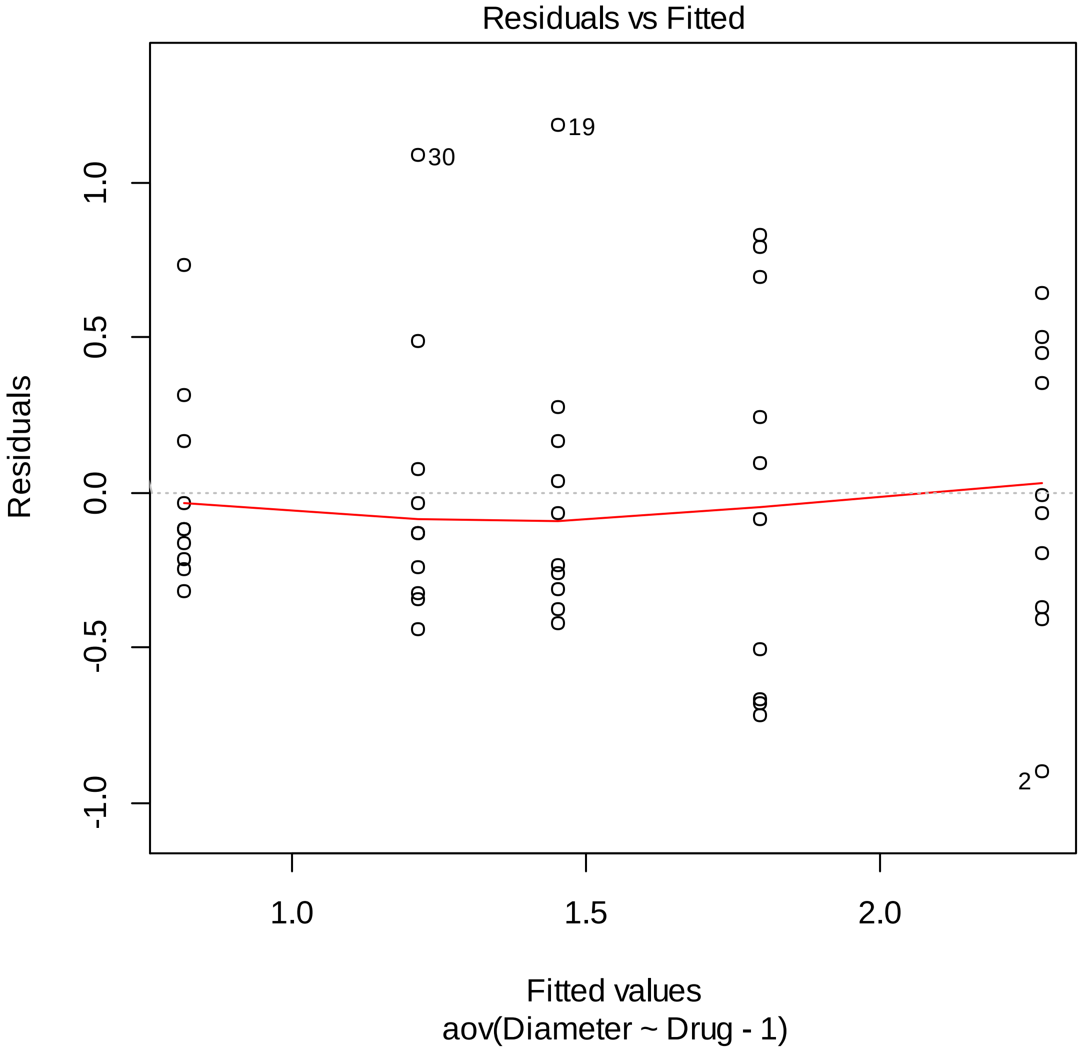
plot(tumorANOVA): Residuals vs. Fitted

Constant variance assumption:

The residuals should have the same vertical spread across the range of the x-axis.

Independence assumption:

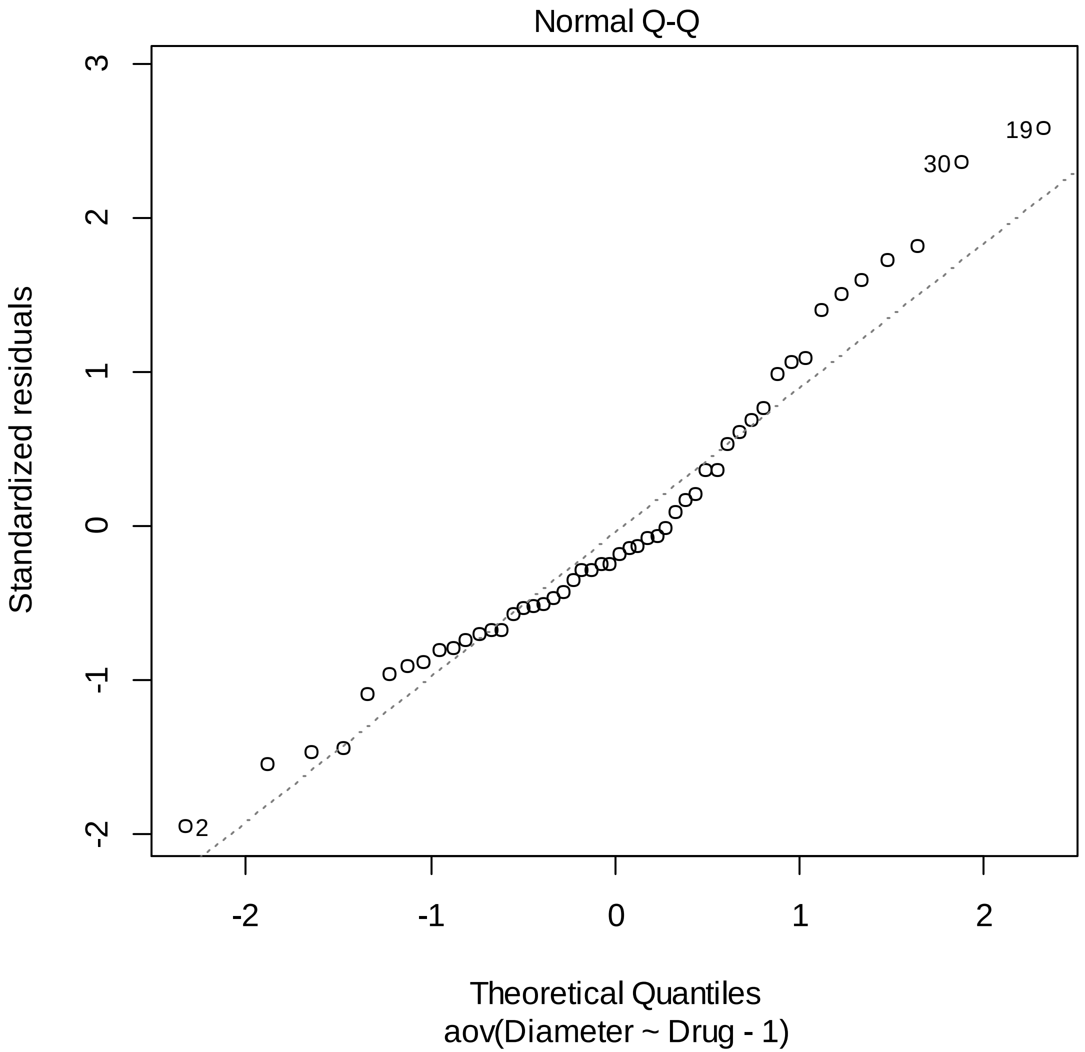
There should be no visible patterns in the residuals (i.e. curved or up-and-down patterns from left-to-right)



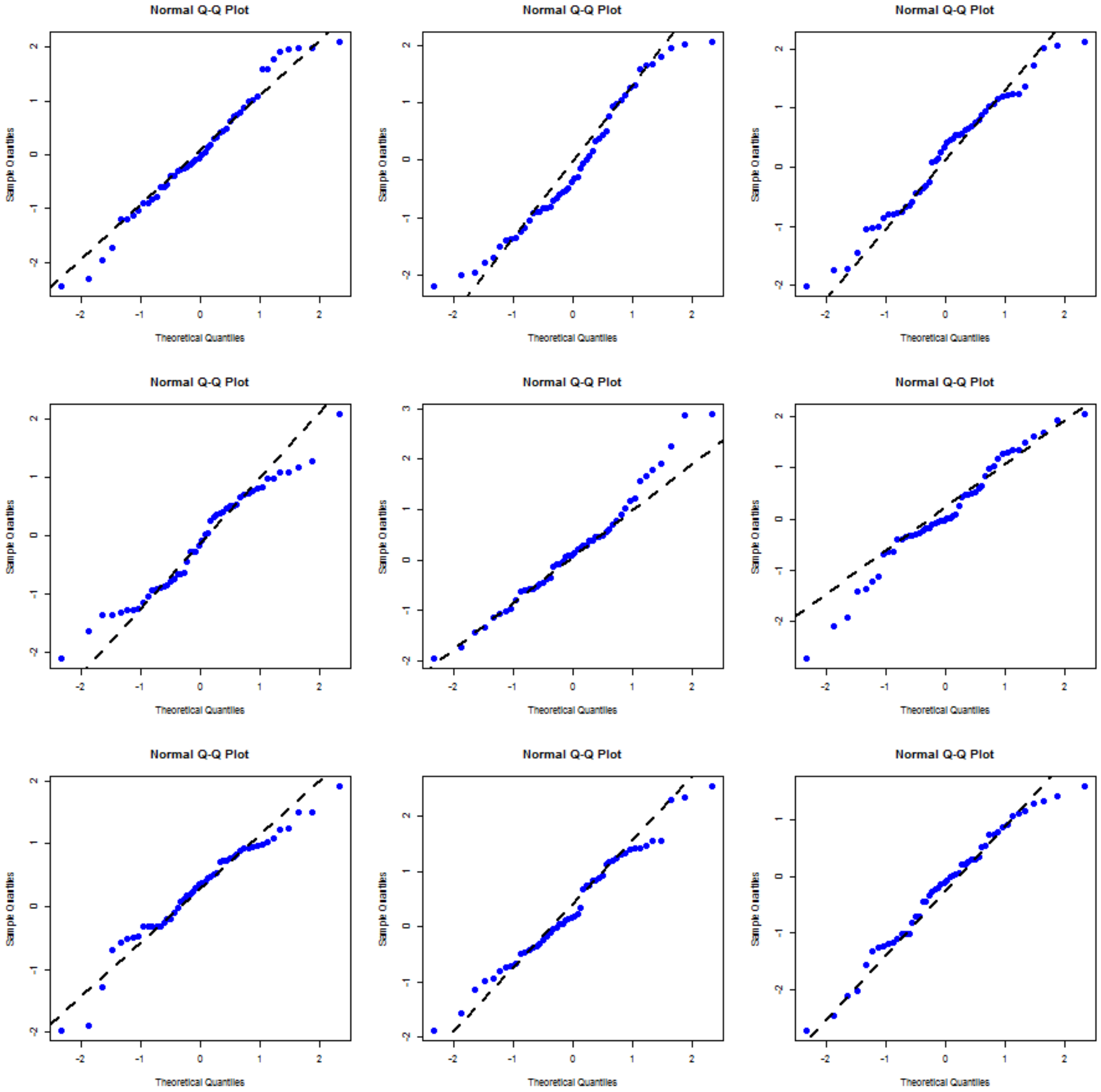
plot(tumorANOVA): Normal Q-Q

Normality assumption:

If the residuals are normally distributed, then the points in the Q-Q plot will be approximately on the dotted line.

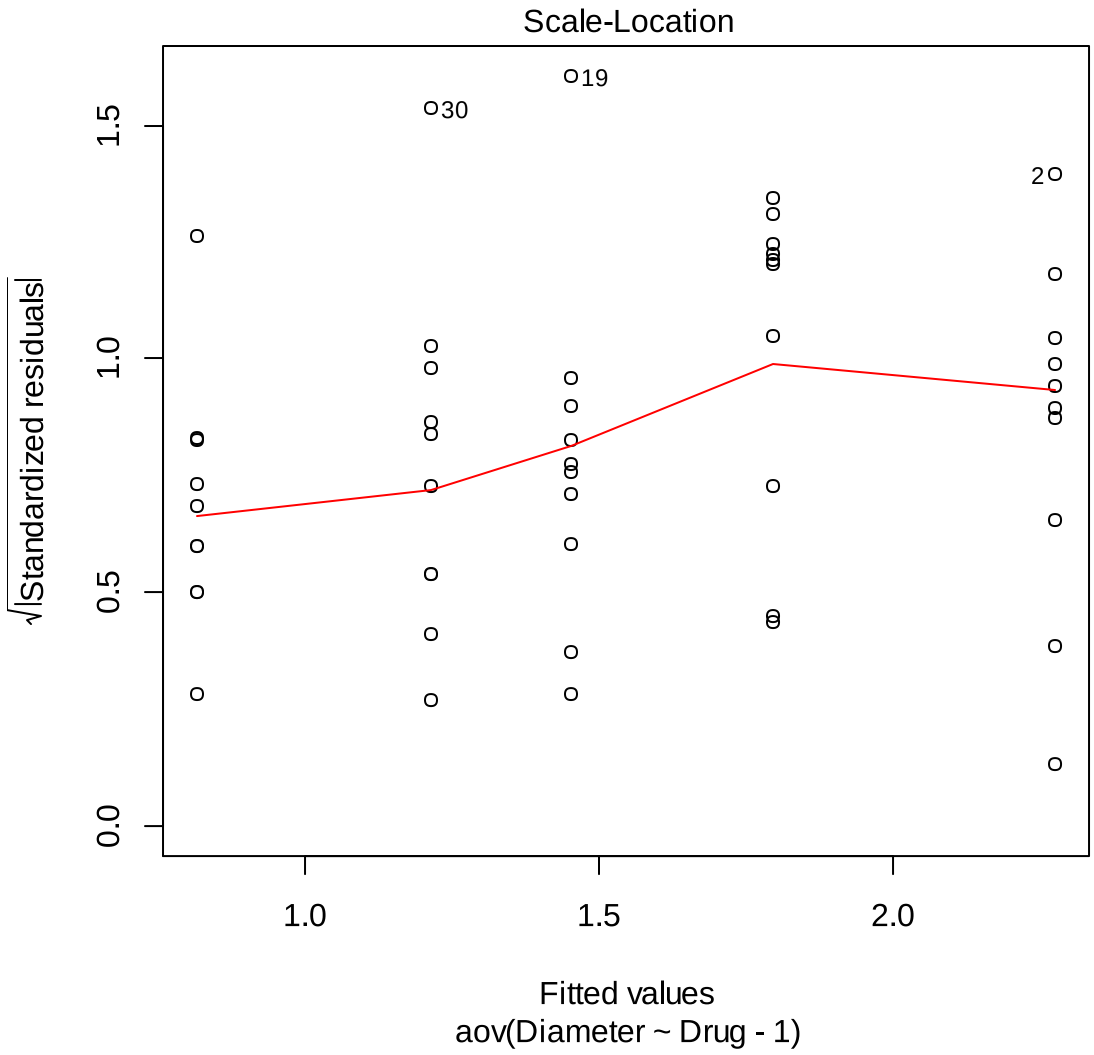


Example Q-Q Plots for Normal Data



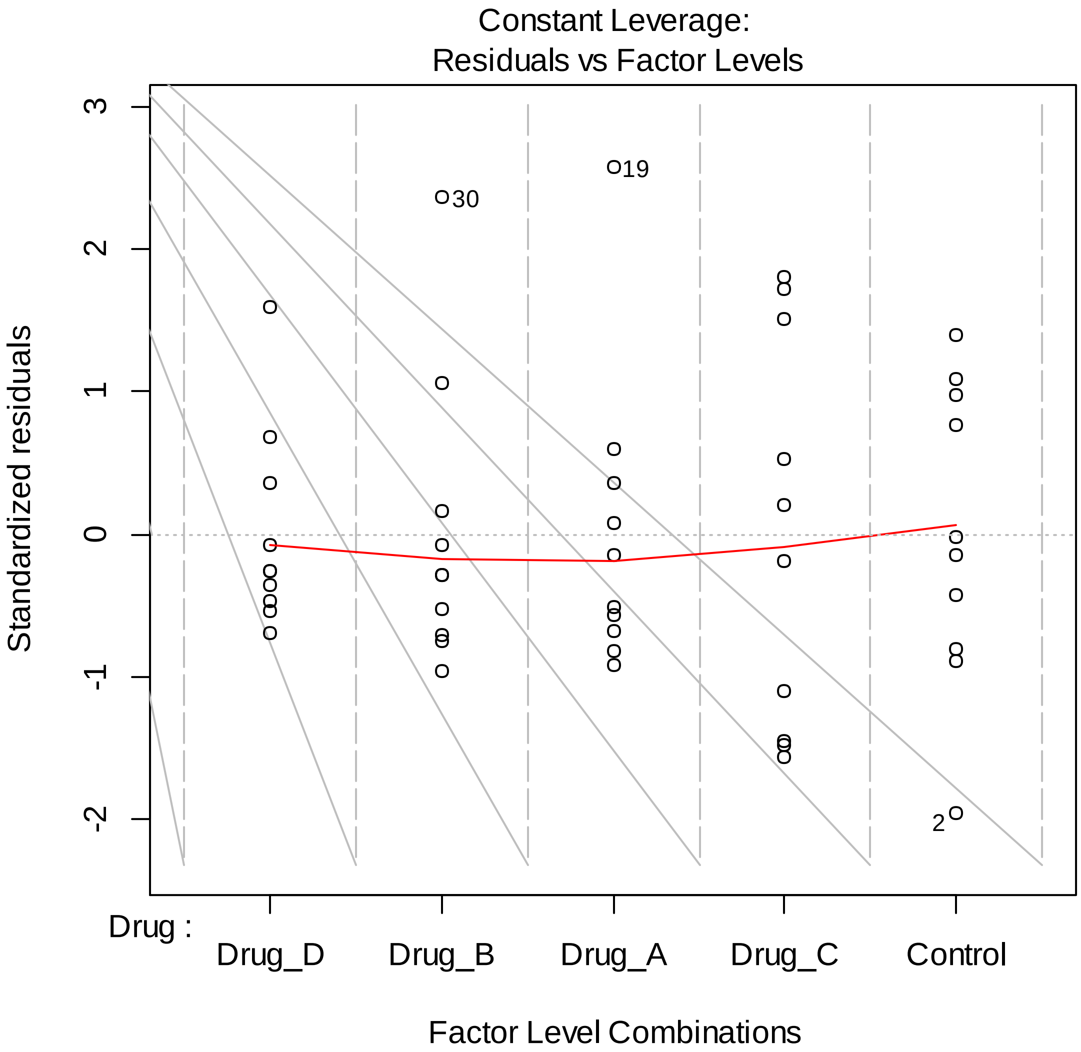
plot(tumorANOVA): Scale-Location

Are any of the residuals too “large”?



plot(tumorANOVA): Leverage

Do any samples have too much impact on the model?

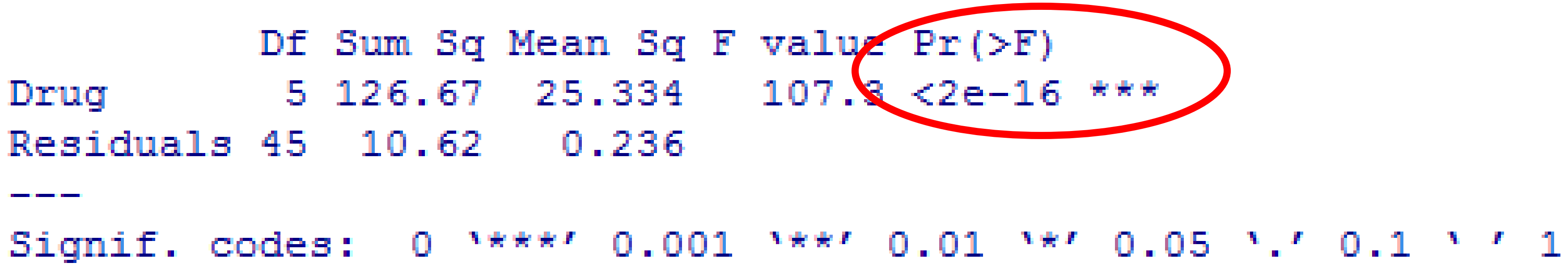


Residuals Are OK. What Next?

- Since the residuals appear to meet the assumptions, we can examine the F-test and corresponding p-value:

Use the summary function with the tumorANOVA object:

```
summary(tumorANOVA)
```



```
          Df Sum Sq Mean Sq F value Pr(>F)    
Drug       5 126.67  25.334   107.3 <2e-16 ***
Residuals 45  10.62   0.236                   
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value is less than 0.05: at least two drugs have different average tumor diameter responses.

What if the Residuals Don’t Meet the Assumptions?

- Variance not constant?
- Non-random pattern in the residuals?
- Residuals don’t follow a normal distribution?
- Contact a friendly statistician….

How do we Compare Two Treatments?

- If the F-test is significant, then at least two treatment means are different. Which two?

- We compare pairs of treatment means using “post-hoc” tests.

- But, we have to take special care when performing these tests to insure that we don’t get false-positive findings

Important Question Prior to Performing Post-Hoc Tests

- Which comparisons are of interest?
- All possible treatment pairs?
  - For 5 treatments, we have 10 possible treatment pairs to compare
- Each treatment to the control?
  - For 4 treatments and a control, there are 4 possible treatment pairs to compare

Controlling for False Positive Findings

- The more tests we perform, the more likely we will find at least one statistically significant comparison, just by chance (not due to a true treatment effect).
- This is called a false positive result
  - We need to insure that we minimize the chance of a false positive finding

- If you desire only pairwise comparisons with the control, then we recommend using **Dunnett's** adjustment.
- If you desire all possible pairwise comparisons, then we recommend using **Tukey's** adjustment.

Dunnett’s Adjustment

- To perform post-hoc tests, we need the multcomp package. We then need to set-up the appropriate contrasts:

Install and load the multcomp package:

```
install.packages("multcomp", dependencies=TRUE)
library(multcomp)
```

Tabulate the number of samples per treatment

```
nDrug = table(tumorNarrow$Drug)
```

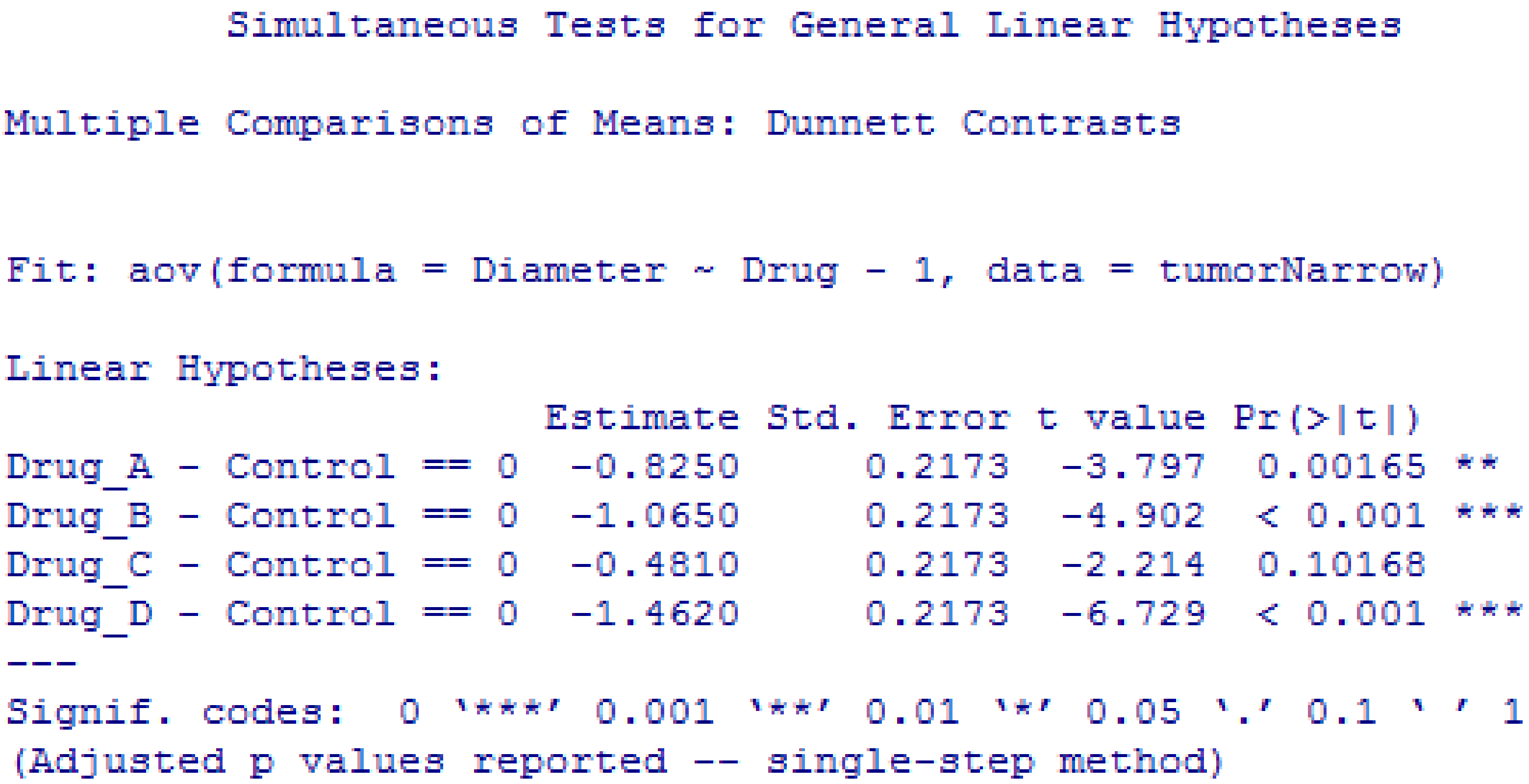
Create the contrast matrix and compute the pairwise comparisons:

```
CMDunnett = contrMat(nDrug)
glhtDunnett = glht(tumorANOVA,linfct=CMDunnett)
```

Dunnett’s Results

Get the summary for the glhtDunnett object:

```
summary(glhtDunnett)
```

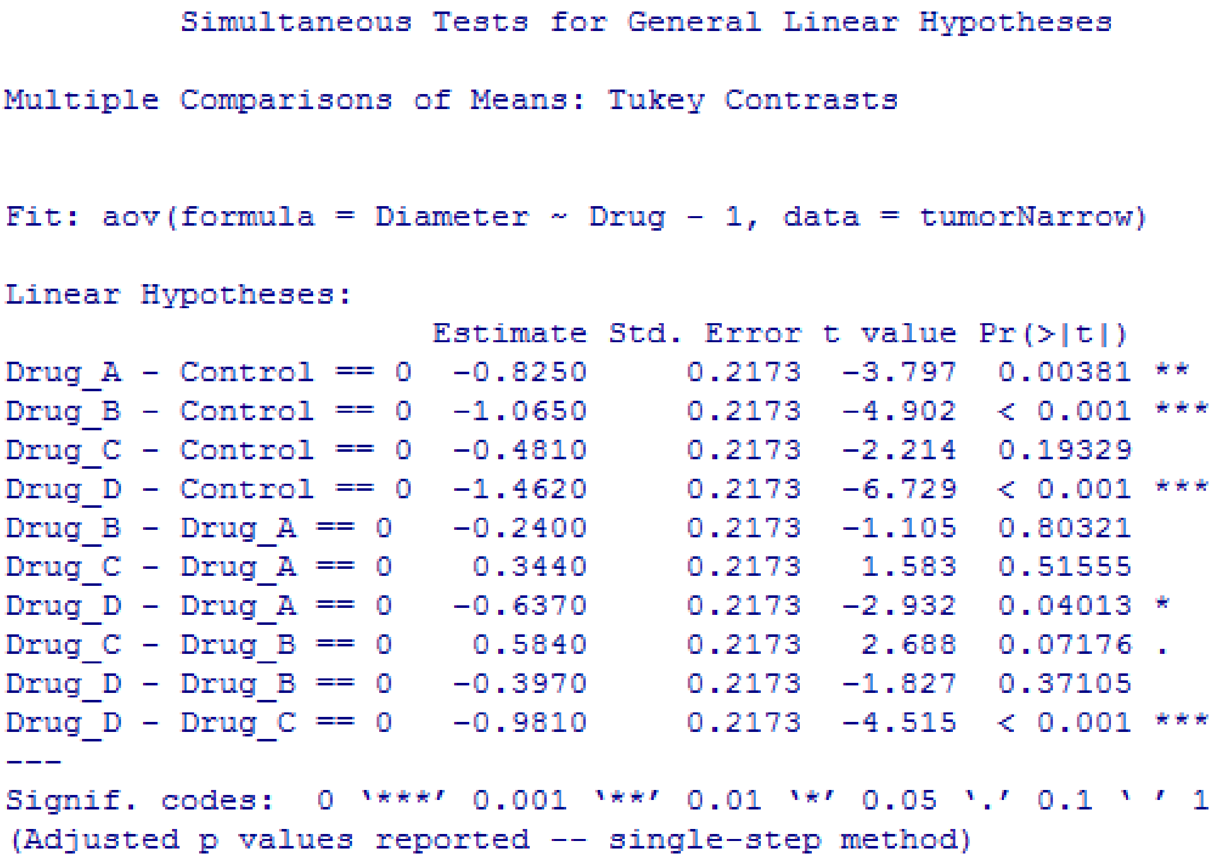


```
         Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Dunnett Contrasts


Fit: aov(formula = Diameter ~ Drug - 1, data = tumorNarrow)

Linear Hypotheses:
                      Estimate Std. Error t value Pr(>|t|)    
Drug_A - Control == 0  -0.8250     0.2173  -3.797  0.00165 ** 
Drug_B - Control == 0  -1.0650     0.2173  -4.902  < 0.001 ***
Drug_C - Control == 0  -0.4810     0.2173  -2.214  0.10168    
Drug_D - Control == 0  -1.4620     0.2173  -6.729  < 0.001 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Adjusted p values reported -- single-step method)
```

Tukey’s Adjustment

Set-up the contrast matrix and compute pairwise tests:

```
CMTukey = contrMat(nDrug, type="Tukey")
glhtTukey = glht(tumorANOVA,linfct=CMTukey)
summary(glhtTukey)
```



```
	 Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts


Fit: aov(formula = Diameter ~ Drug - 1, data = tumorNarrow)

Linear Hypotheses:
                      Estimate Std. Error t value Pr(>|t|)    
Drug_A - Control == 0  -0.8250     0.2173  -3.797  0.00381 ** 
Drug_B - Control == 0  -1.0650     0.2173  -4.902  < 0.001 ***
Drug_C - Control == 0  -0.4810     0.2173  -2.214  0.19329    
Drug_D - Control == 0  -1.4620     0.2173  -6.729  < 0.001 ***
Drug_B - Drug_A == 0   -0.2400     0.2173  -1.105  0.80321    
Drug_C - Drug_A == 0    0.3440     0.2173   1.583  0.51555    
Drug_D - Drug_A == 0   -0.6370     0.2173  -2.932  0.04013 *  
Drug_C - Drug_B == 0    0.5840     0.2173   2.688  0.07176 .  
Drug_D - Drug_B == 0   -0.3970     0.2173  -1.827  0.37105    
Drug_D - Drug_C == 0   -0.9810     0.2173  -4.515  < 0.001 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Adjusted p values reported -- single-step method)
```

Upcoming Sessions

- Part 3: Comparing Groups (2)
  - Fixed and random effects, how to model data with mixed (fixed and random) effects, repeated measures data, visualization
- Part 4: Covariance Structures in Mixed Models and Dimension Reduction and Classification
  - Principal component analysis (PCA), partial least squares (PLS), recursive partitioning (RPart), and random forests (RF)