Is ANOVA the same as Linear Regression?

07-15-2021

Basically yes!

ANOVA and Linear Regression are both Linear Models

$$Y = a + bX$$

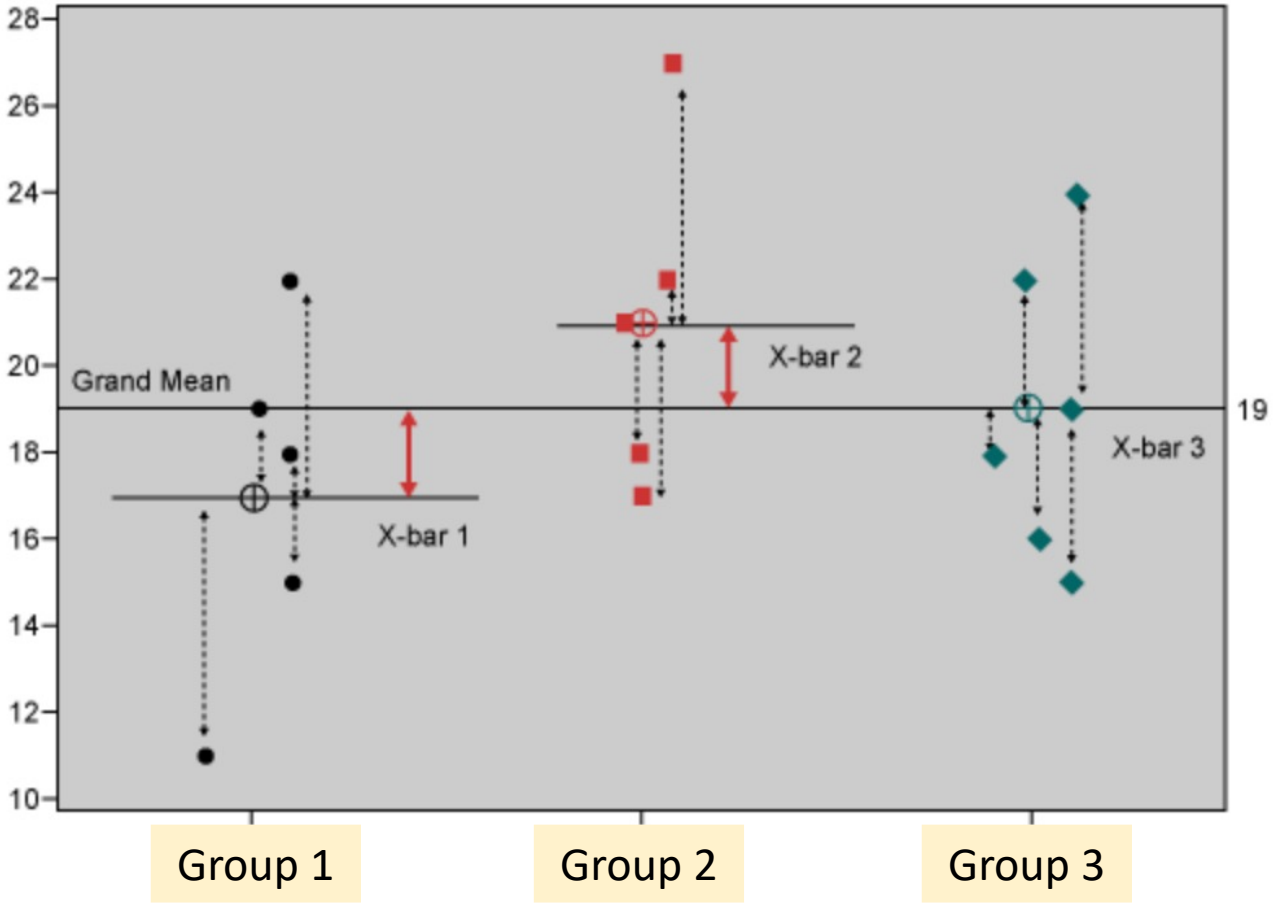
Linear Models describe a continuous ‘response’ variable as a function of one or more ‘predictor’ variables!

What is an ANOVA?

Statistical analysis when assessing for differences between group means on a continuous measurement

Independent variable(s) are qualitative

Dependent variable is quantitative

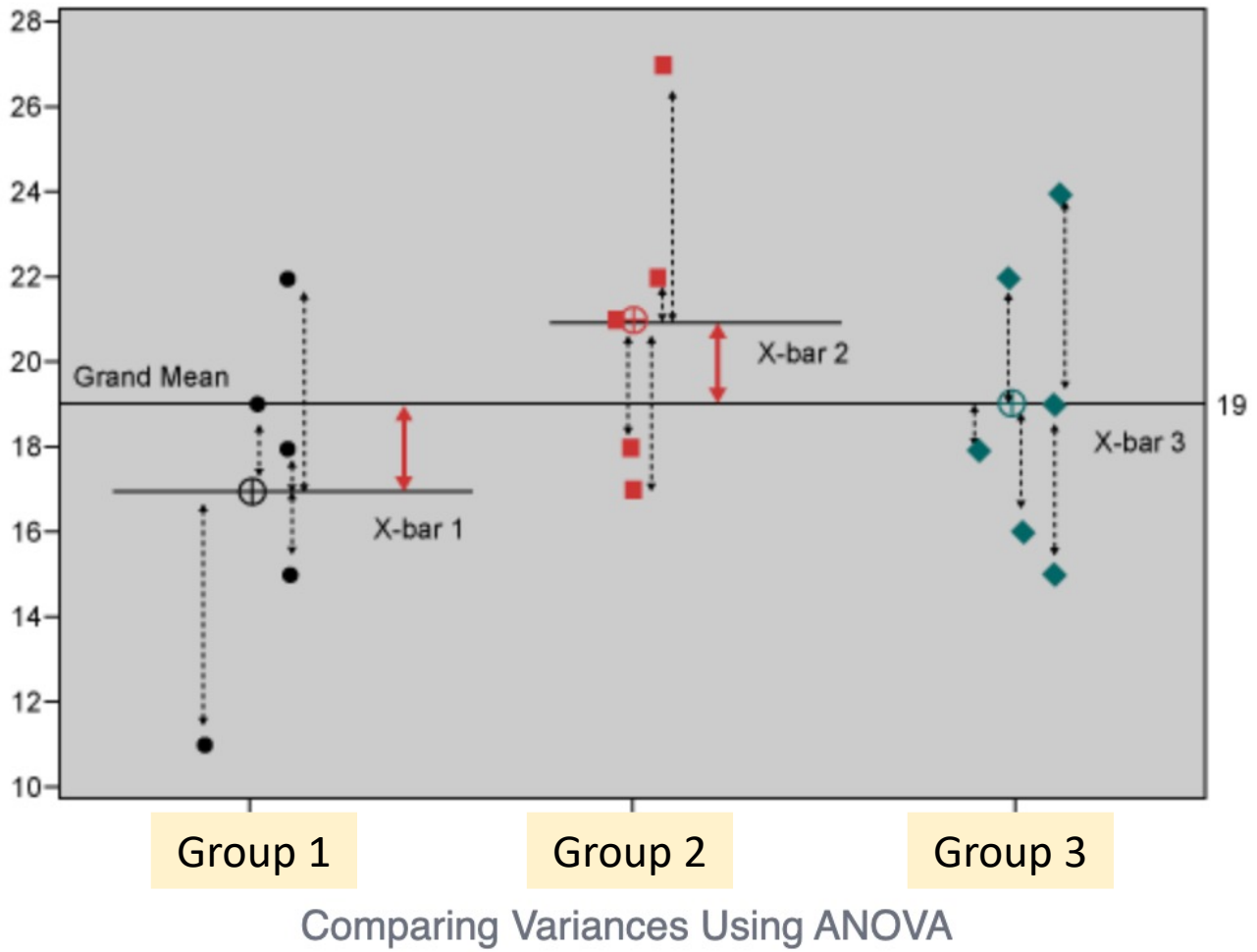


Comparing Variances Using ANOVA

What is an ANOVA?

Compares 2 types of variance:
1) ‘between’ groups
2) ‘within’ each group

Assumption:
If population means are different then the variance ‘within’ the samples must be small compared to the variance ‘between’ samples



Comparing Variances Using ANOVA

What is an ANOVA?

Summary Table for the One-way ANOVA

Summary ANOVA

Source	Sum of Squares	Degrees of Freedom	Variance Estimate (Mean Square)	*F* Ratio
Between	SS_B	$K-1$	$MS_B = \frac{SS_B}{K-1}$	$\frac{MS_B}{MS_W}$
Within	SS_W	$N-K$	$MS_W = \frac{SS_W}{N-K}$	
Total	$SS_T = SS_B + SS_W$	$N-1$		

Knowing that K (Groups) = 5 and N (Total Sample Size) = 50 (n = 10 for each group)…

Table 1

Analysis of Variance for Number of Words Recalled

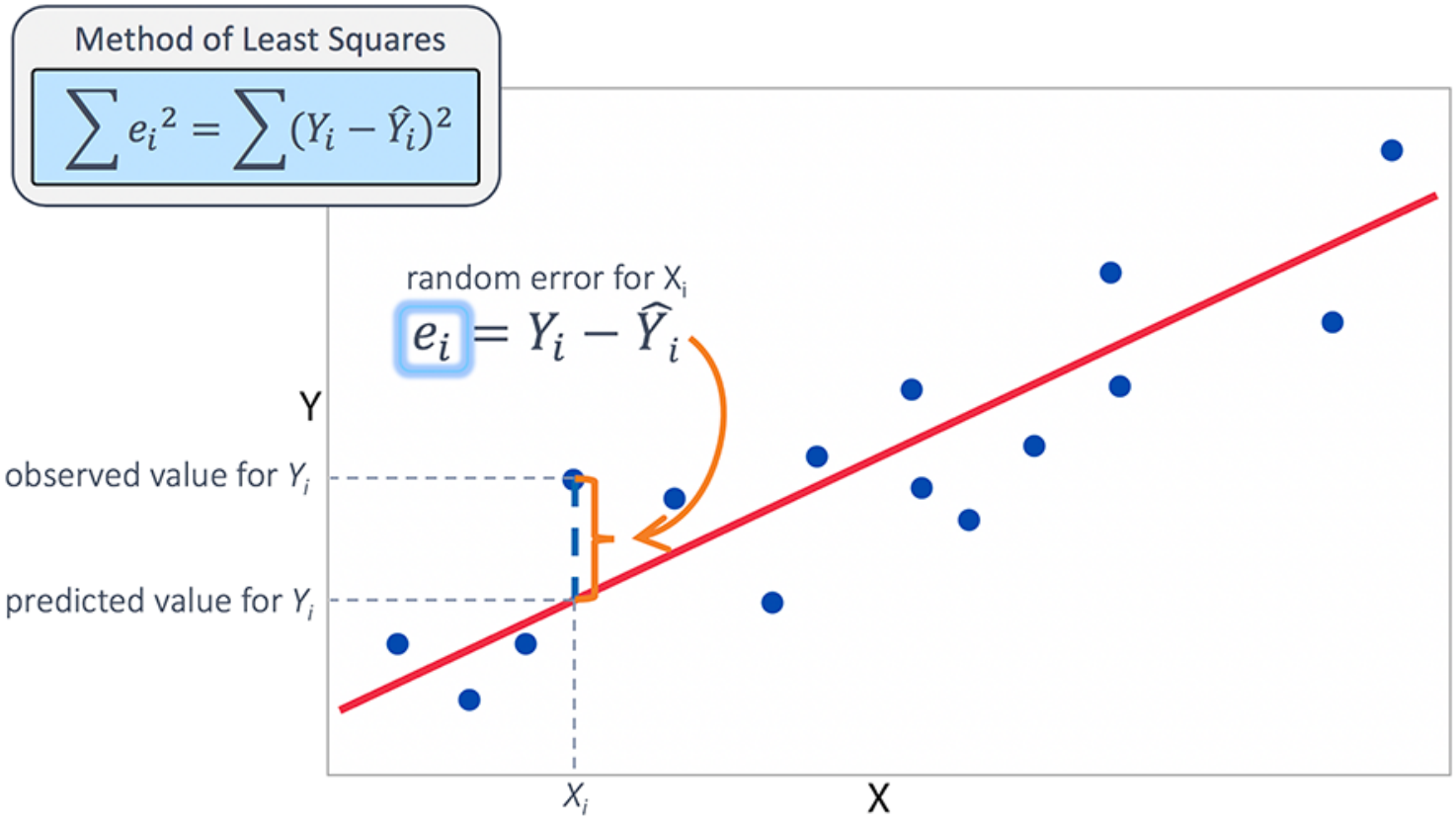
Source	*SS*	*df*	*MS*	*F*	F_{CV}
Between	351.52	4	87.88	9.08*	2.61
Within	435.30	45	9.67		
Total	786.82	49			

* $p < .05$

What is Linear Regression?

Independent and dependent variables are quantitative

Finds the line of 'best fit' through your data by looking for the regression coefficient that minimizes the total error of the model



Let’s now compare ANOVA and Linear Regression by answering a biological question….

How does RNAi and drug treatment affect pancreatic gene expression of *INS in vitro*?

Human pancreatic cells were treated with either…

1. RNAi negative control
2. RNAi (against *INS)*
3. Metformin
4. RNAi + Metformin

INS gene expression was measured

Let’s make the dataset!

For ANOVA model…..

Data is coded like so:

1. Factor/independent variables
    a. RNAi
        i. Yes = 1 ’treated’
        ii. No = 0 ‘untreated’
    b. Metformin
        i. Yes = 1
        ii. No = 0

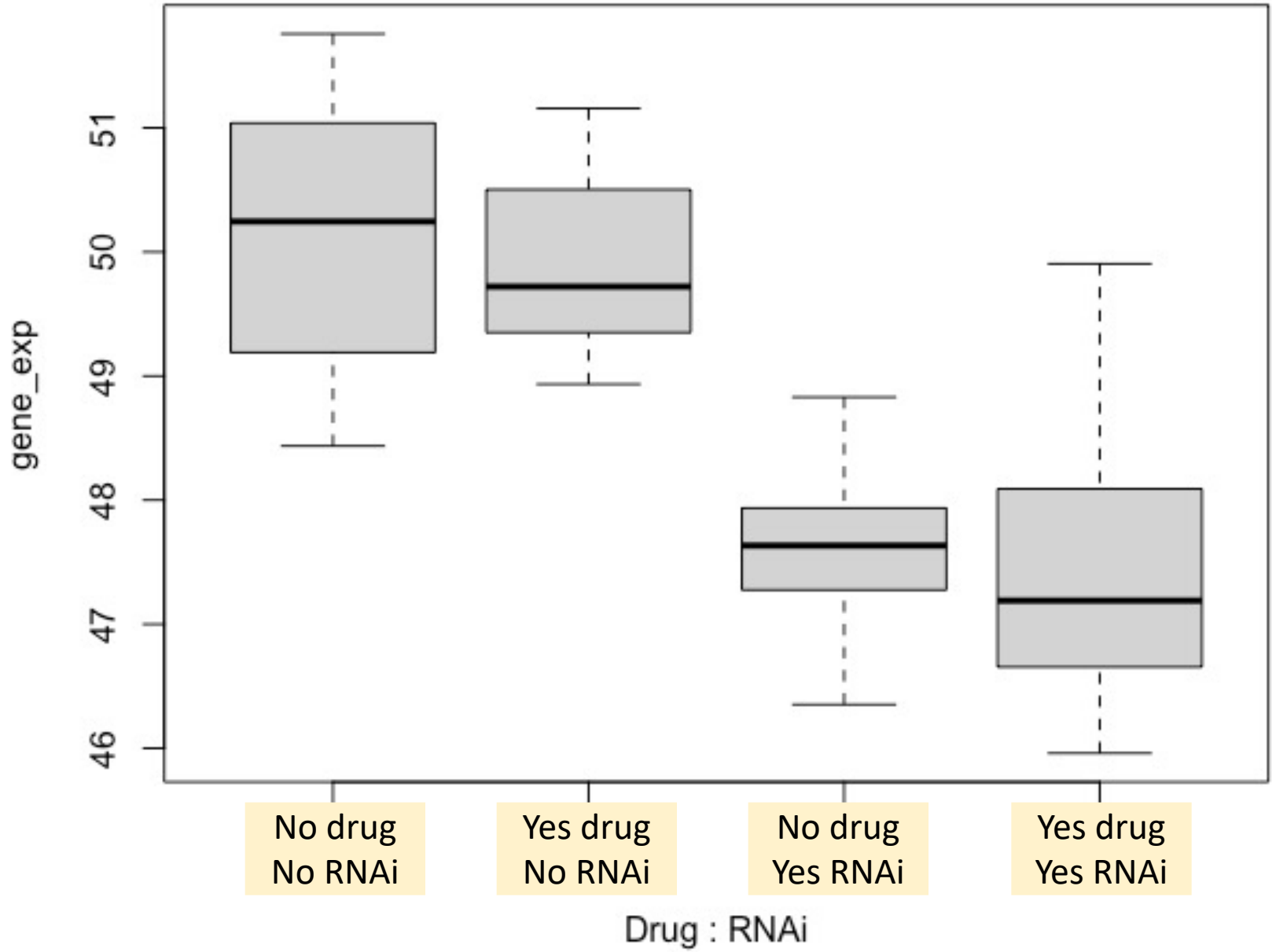
2. Response/dependent variable
    a. *INS* gene expression

```r
# Make fake data for ANOVA
set.seed(0)
data = data.frame(Sample=1:50,
                  gene_exp = c( rnorm(25, 47.5), # treated with RNAi
                                rnorm(25, 50)), # RNAi-negative control
                  RNAi = c( rep(1,25), # RNAi treated are first 25
                            rep(0,25)),
                  Drug = rep(c(0,1),25)) # drug treatment is not associated with a drop in expression

data$RNAi = factor(data$RNAi)
data$Drug = factor(data$Drug)
```

```
> print(data)
   Sample gene_exp RNAi Drug
1       1 48.76295    1    0
2       2 47.17377    1    1
3       3 48.82980    1    0
4       4 48.77243    1    1
5       5 47.91464    1    0
6       6 45.96005    1    1
7       7 46.57143    1    0
8       8 47.20528    1    1
9       9 47.49423    1    0
10     10 49.90465    1    1
```

```
> str(data)
'data.frame':   50 obs. of  4 variables:
 $ Sample  : int  1 2 3 4 5 6 7 8 9 10 ...
 $ gene_exp: num  48.8 47.2 48.8 48.8 47.9 ...
 $ RNAi    : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 2 ...
 $ Drug    : Factor w/ 2 levels "0","1": 1 2 1 2 1 2 1 2 1 2 ...
```

```
boxplot(gene_exp ~ Drug + RNAi, data)
```



Hypotheses: $\mathbf{H_0}$: $\mu_1 = \mu_2 = \ldots = \mu_k$ or $\mathbf{H_A}$: $\mu_1 \neq \mu_2 \neq \ldots \neq \mu_k$

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{12} X_1 X_2 + \in_i$$

(response variable)
***INS* gene expression level**

(intercept)
Base level of measurement in the negative control

(predictor)
RNAi transfected = 1
RNAi transfected = 0

(predictor)
Drug administered = 1
Drug not administered = 0

(interaction)
RNAi transfected = 1
RNAi transfected = 0
Drug administered = 1
Drug not administered = 0

Hypotheses: $\mathbf{H_0}$: $\mu_1 = \mu_2 = \ldots = \mu_k$ or $\mathbf{H_A}$: $\mu_1 \neq \mu_2 \neq \ldots \neq \mu_k$

1way ANOVA

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{12} X_1 X_2 + \in_i$$

(response variable)
***INS* gene expression level**

(intercept)
Base level of measurement in the negative control

(predictor)
RNAi transfected = 1
RNAi transfected = 0

(predictor)
Drug administered = 1
Drug not administered = 0

(interaction)
RNAi transfected = 1
RNAi transfected = 0
Drug administered = 1
Drug not administered = 0

Hypotheses: **H_0:** $\mu_1 = \mu_2 = \ldots = \mu_k$ or **H_A:** $\mu_1 \neq \mu_2 \neq \ldots \neq \mu_k$

2way ANOVA

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{12} X_1 X_2 + \in_i$$

(response variable)
***INS* gene expression level**

(intercept)
Base level of measurement in the negative control

(predictor)
RNAi transfected = 1
RNAi transfected = 0

(predictor)
Drug administered = 1
Drug not administered = 0

(interaction)
RNAi transfected = 1
RNAi transfected = 0
Drug administered = 1
Drug not administered = 0

Hypotheses: $\mathbf{H_0}$: $\mu_1 = \mu_2 = \ldots = \mu_k$ or $\mathbf{H_A}$: $\mu_1 \neq \mu_2 \neq \ldots \neq \mu_k$

2way ANOVA

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{12} X_1 X_2 + \in_i$$

(response variable)
***INS* gene expression level**

(intercept)
Base level of measurement in the negative control

(predictor)
RNAi transfected = 1
RNAi transfected = 0

(predictor)
Drug administered = 1
Drug not administered = 0

(interaction)
RNAi transfected = 1
RNAi transfected = 0
Drug administered = 1
Drug not administered = 0

```r
# Analyze data with ANOVA model
one.way <- aov(gene_exp ~ Drug, data)
summary(one.way)
```

```
> summary(one.way)
             Df Sum Sq Mean Sq F value Pr(>F)
Drug          1    0.2  0.1991   0.082  0.776
Residuals    48  117.1  2.4389
```

```r
# Analyze data with ANOVA model
one.way <- aov(gene_exp ~ Drug, data)
summary(one.way)

two.way <- aov(gene_exp ~ Drug + RNAi, data)
summary(two.way)
```

```
> summary(two.way)
             Df Sum Sq Mean Sq F value   Pr(>F)
Drug          1   0.20    0.20   0.231    0.633
RNAi          1  76.57   76.57  88.876 2.07e-12 ***
Residuals    47  40.49    0.86
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```r
# Analyze data with ANOVA model
one.way <- aov(gene_exp ~ Drug, data)
summary(one.way)

two.way <- aov(gene_exp ~ Drug + RNAi, data)
summary(two.way)

two.way_interaction <- aov(gene_exp ~ Drug + RNAi + Drug*RNAi, data)
summary(two.way_interaction)
```

```
> summary(two.way_interaction)
             Df Sum Sq Mean Sq F value   Pr(>F)
Drug          1   0.20    0.20   0.226    0.637
RNAi          1  76.57   76.57  86.985 3.56e-12 ***
Drug:RNAi     1   0.00    0.00   0.000    0.998
Residuals    46  40.49    0.88
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

How do you decide which ANOVA model to use?

We want to use the ‘best-fit’ model, the model that best explains the variation in the dependent variable!

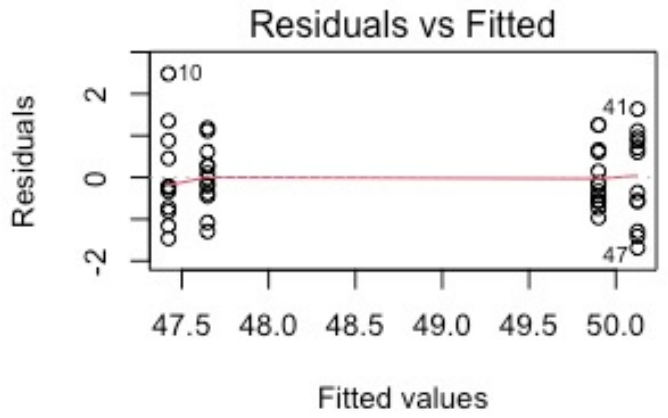
The Akaike (Ah-KYE-EE-KAY) information criterion (AIC) is a test for model fit.

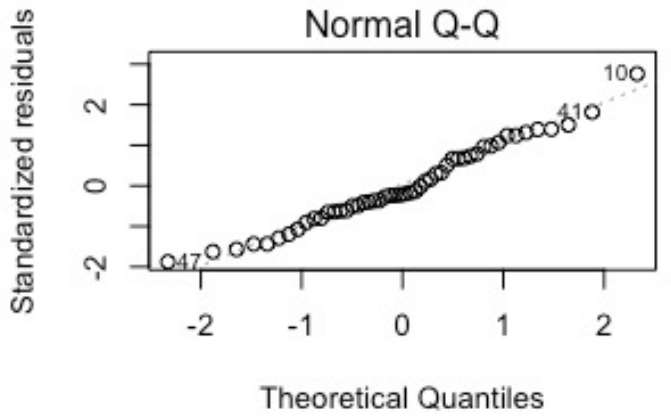
```
#Find the best-fit ANOVA model
model.set <- list(one.way, two.way, two.way_interaction)
model.names <- c("one.way", "two.way", "interaction")

aictab(model.set, modnames = model.names)
```

```
Model selection based on AICc:

            K   AICc Delta_AICc AICcWt Cum.Wt     LL
two.way     4 140.24       0.00   0.78   0.78 -65.68
interaction 5 142.71       2.47   0.22   1.00 -65.68
one.way     3 190.95      50.71   0.00   1.00 -92.22
```

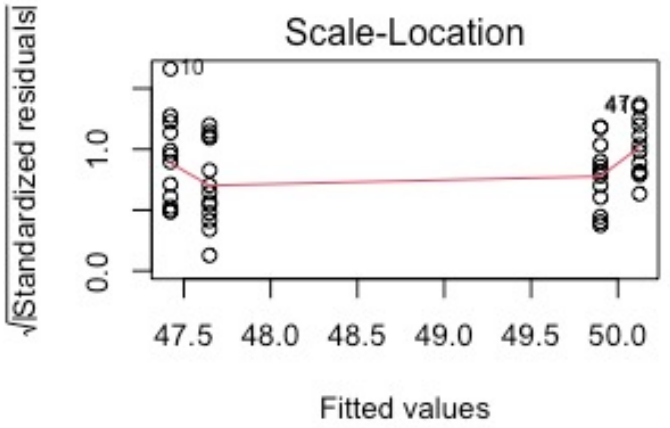
Now check whether the model fits the assumption of homoscedasticity

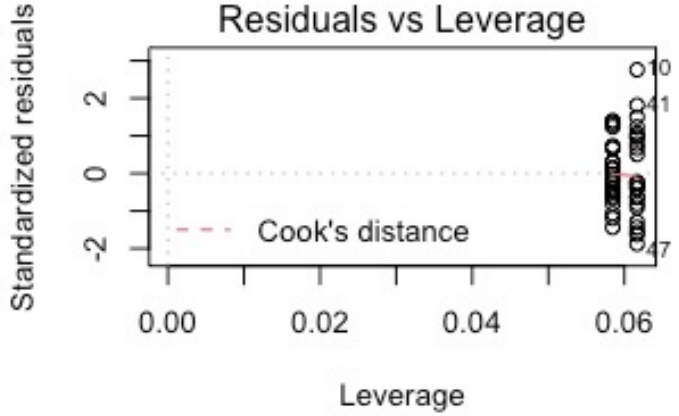
```
plot(two.way)
```





Diagnostic plots show the unexplained variance (residuals) across the range of the observed data

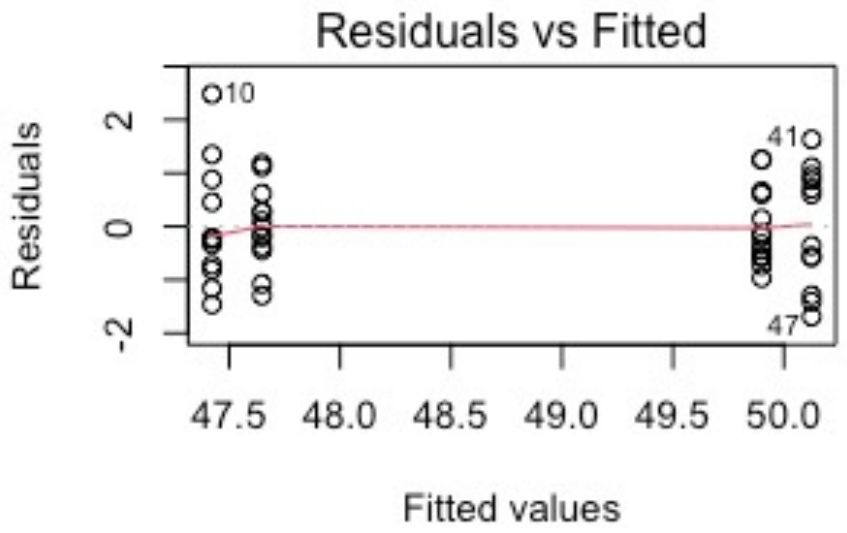


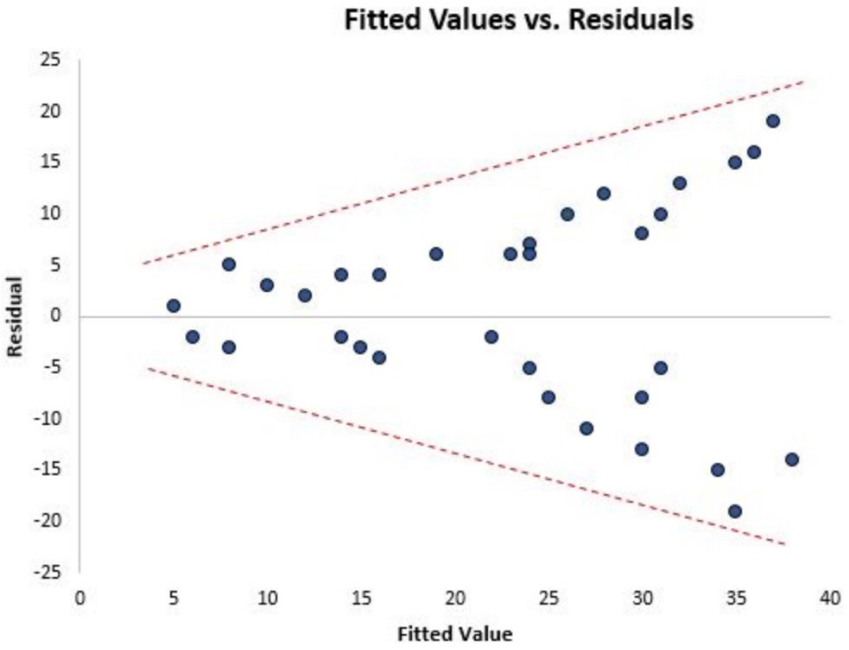


https://data.library.virginia.edu/diagnostic-plots/

Now check whether the model fits the assumption of homoscedasticity

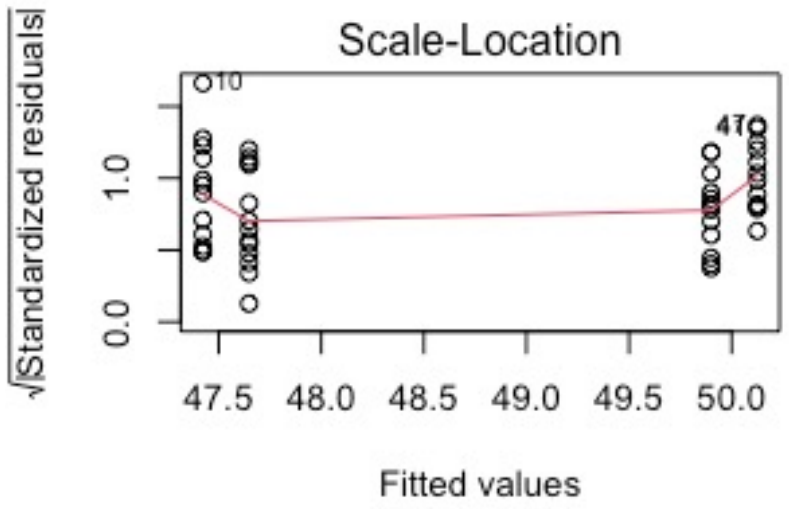
‘Residuals vs Fitted’
Shows if the variability of the observations differs across groups because all observations in the same group get the same fitted value

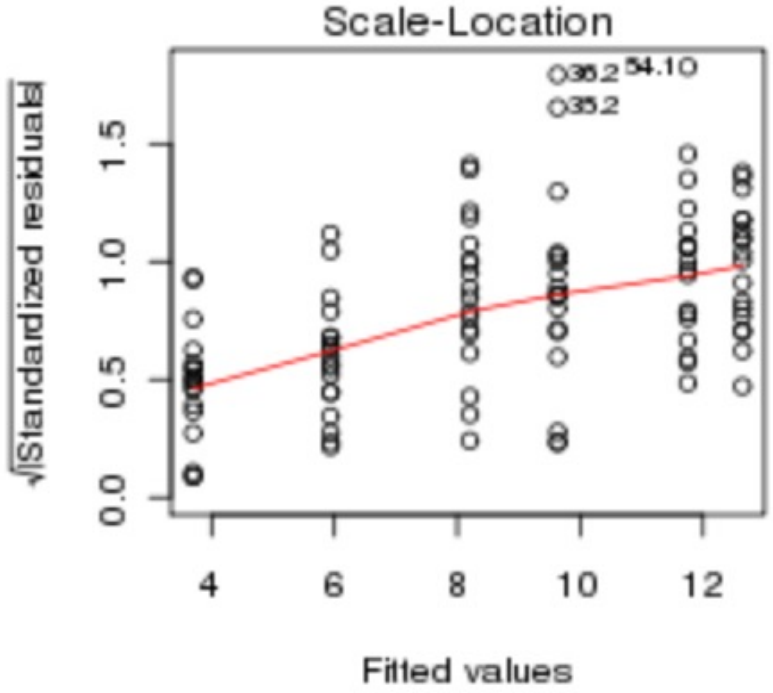




Now check whether the model fits the assumption of homoscedasticity

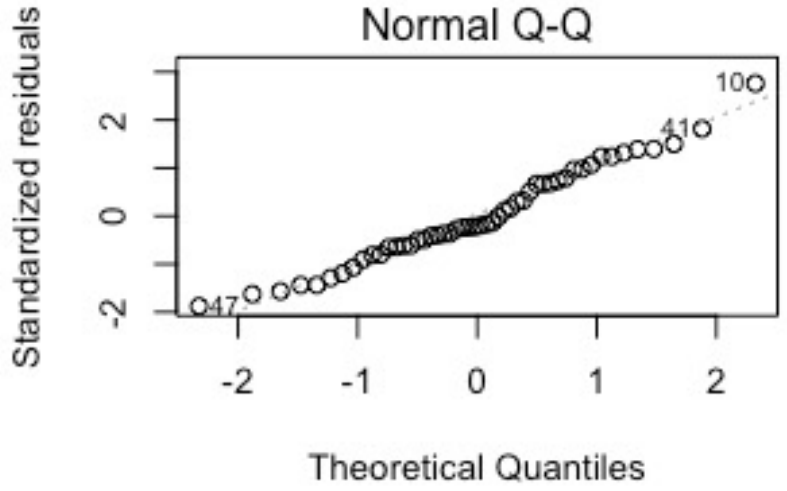
‘Scale-location’
Again, you want to assess whether the groups have somewhat similar or noticeably different amounts of variability

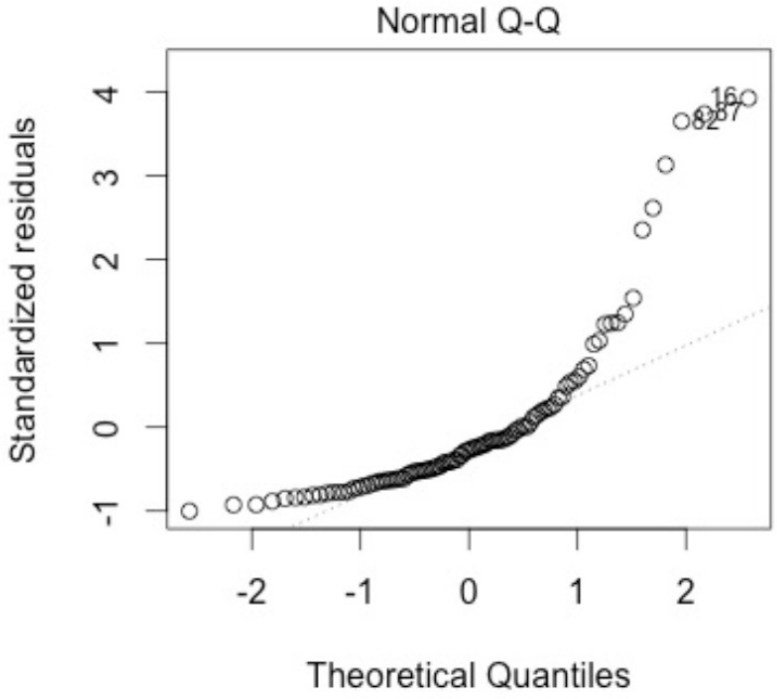




Now check whether the model fits the assumption of homoscedasticity

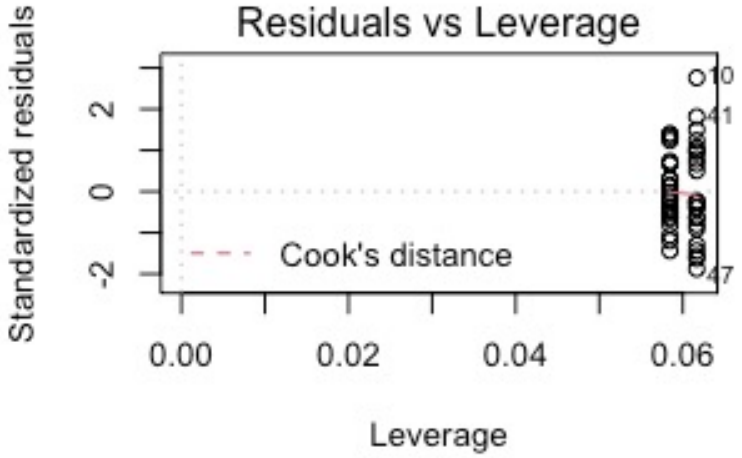
‘Normal Q-Q’
Shows if residuals are normally distributed

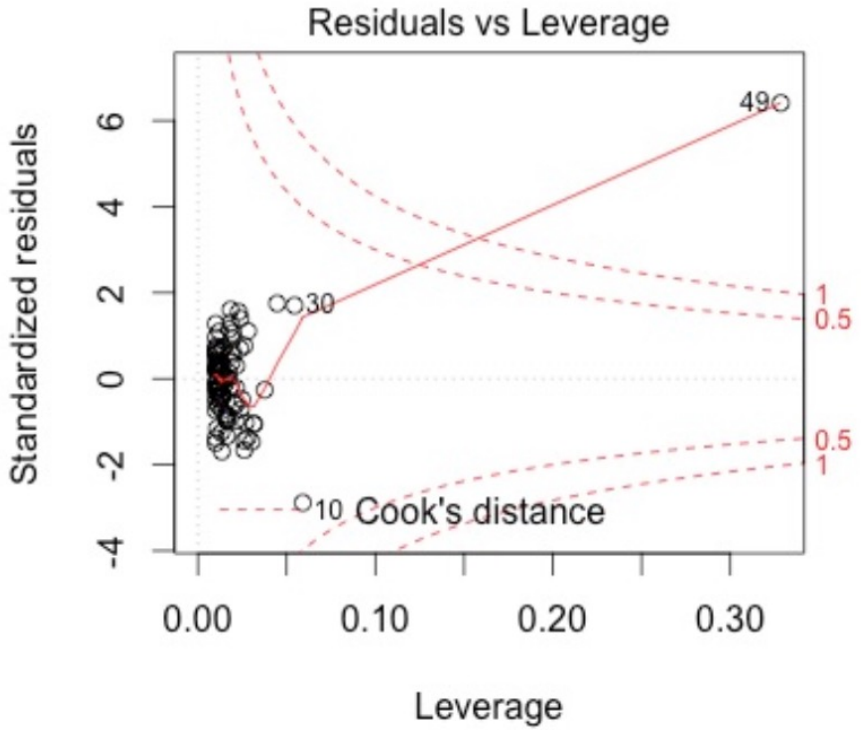




Now check whether the model fits the assumption of homoscedasticity

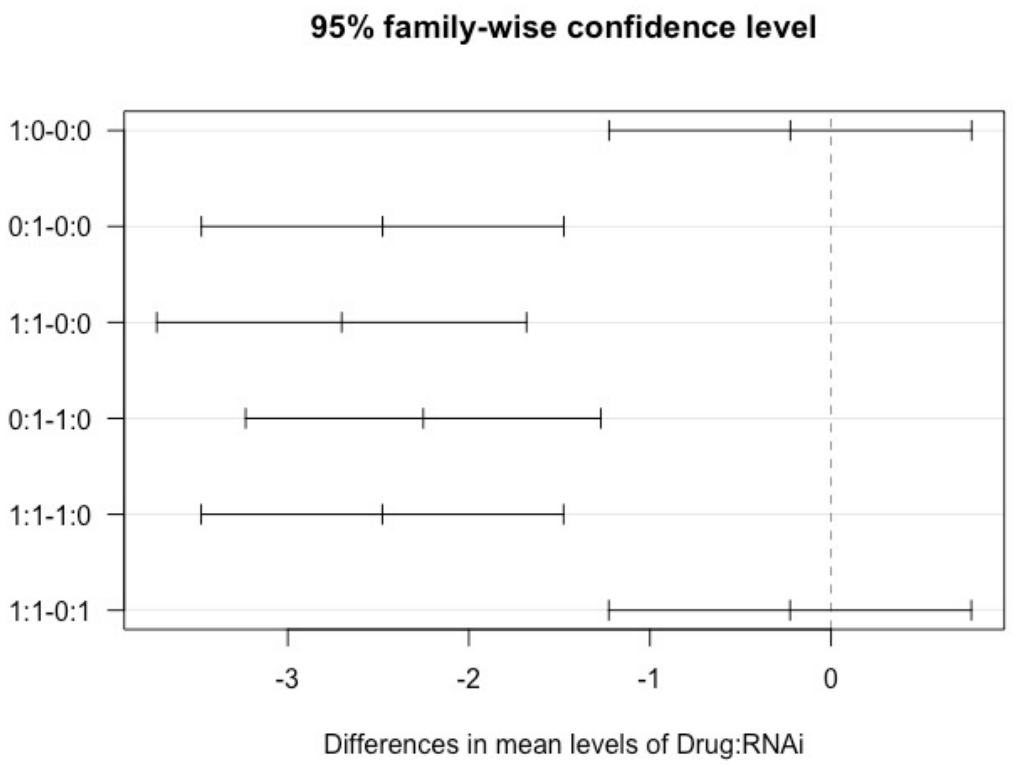
‘Residuals vs Leverage’
Helps identify
influential data points
on the model





What are the differences if any among our groups?

```
# Post hoc test
tukey.plot.test <-TukeyHSD(two.way_interaction)
plot(tukey.plot.test, las = 1)
tukey.plot.test
```




```
> tukey.plot.test
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = gene_exp ~ Drug + RNAi + Drug * RNAi, data = data)

$Drug
          diff        lwr       upr     p adj
1-0 -0.1262047 -0.6603786 0.4079693 0.6366319

$RNAi
         diff       lwr       upr p adj
1-0 -2.473072 -3.007246 -1.938898     0

$`Drug:RNAi`
               diff       lwr        upr     p adj
1:0-0:0 -0.2244977 -1.225655  0.7766594 0.9322469
0:1-0:0 -2.4762472 -3.477404 -1.4750900 0.0000002
1:1-0:0 -2.7023216 -3.723306 -1.6813376 0.0000000
0:1-1:0 -2.2517494 -3.232679 -1.2708198 0.0000011
1:1-1:0 -2.4778239 -3.478981 -1.4766667 0.0000002
1:1-0:1 -0.2260744 -1.227232  0.7750827 0.9309432
```

Let’s make the dataset!

For Linear Regression model…..

Data is coded like so:

1. Predictors/independent variables
    a. RNAi
        i. Dosage: 0 – 300 nM
    b. Metformin
        i. Dosage: 0 – 500 mg

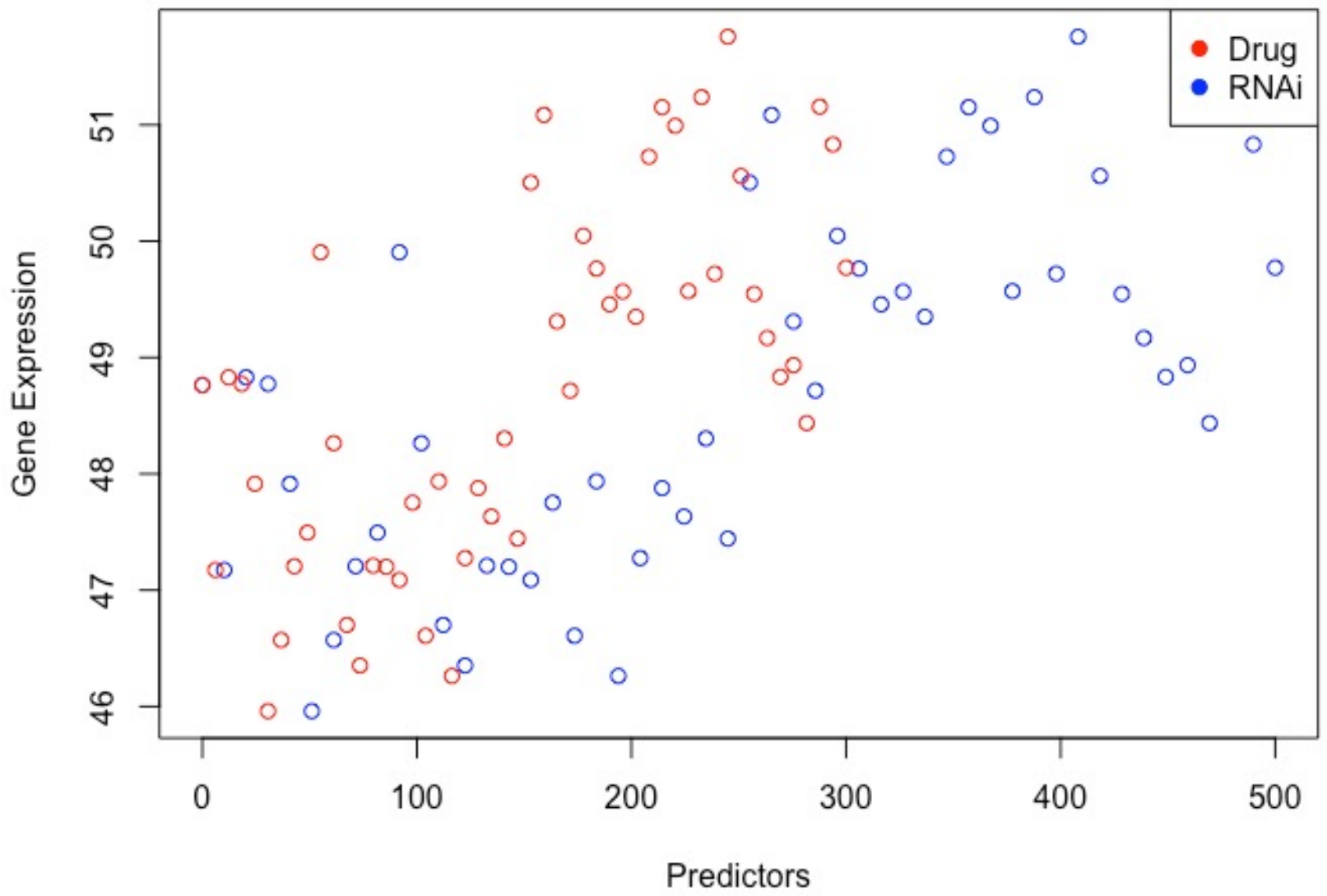
2. Response/dependent variable
    a. *INS* gene expression

```r
# Make fake data for Linear Regression
set.seed(0)
data2 = data.frame(Sample=1:50,
                   gene_exp = c( rnorm(25, 47.5), # treated with RNAi
                                 rnorm(25, 50)), # RNAi-negative control
                   RNAi = seq(from = 0, to = 300, length.out = 50), # RNAi treatment dosage
                   Drug = seq(from = 0, to = 500, length.out = 50)) # Drug treatment dosage
```

```
> print(data2)
   Sample gene_exp       RNAi      Drug
1       1 48.76295   0.000000   0.00000
2       2 47.17377   6.122449  10.20408
3       3 48.82980  12.244898  20.40816
4       4 48.77243  18.367347  30.61224
5       5 47.91464  24.489796  40.81633
6       6 45.96005  30.612245  51.02041
7       7 46.57143  36.734694  61.22449
8       8 47.20528  42.857143  71.42857
9       9 47.49423  48.979592  81.63265
10     10 49.90465  55.102041  91.83673
```

```
> str(data2)
'data.frame':   50 obs. of  4 variables:
 $ Sample  : int  1 2 3 4 5 6 7 8 9 10 ...
 $ gene_exp: num  48.8 47.2 48.8 48.8 47.9 ...
 $ RNAi    : num  0 6.12 12.24 18.37 24.49 ...
 $ Drug    : num  0 10.2 20.4 30.6 40.8 ...
```

```r
plot(data2$Drug, data2$gene_exp, col = 'blue', xlab = 'Predictors', ylab = 'Gene Expression')
points(data2$RNAi, data2$gene_exp, col = 'red')
legend("topright", c("Drug", "RNAi"), pch=c(19,19), col = c("blue", "red"))
```



```r
# Analyze data with Linear Regression model
Data2_lm1 <- lm(gene_exp ~ RNAi, data)
summary(Data2_lm1)
```

```
> summary(Data2_lm1)

Call:
lm(formula = gene_exp ~ RNAi, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-1.5799 -0.6308 -0.2395  0.7225  2.3647

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  50.0079     0.1851 270.124  < 2e-16 ***
RNAi1        -2.4680     0.2618  -9.427 1.69e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9256 on 48 degrees of freedom
Multiple R-squared:  0.6493,    Adjusted R-squared:  0.642
F-statistic: 88.86 on 1 and 48 DF,  p-value: 1.687e-12
```

```r
# Analyze data with Linear Regression model
Data2_lm1 <- lm(gene_exp ~ RNAi, data)
summary(Data2_lm1)

Data2_lm2 <- lm(gene_exp ~ RNAi + Drug, data)
summary(Data2_lm2)
```

```
> summary(Data2_lm2)

Call:
lm(formula = gene_exp ~ RNAi + Drug, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-1.6889 -0.5730 -0.1922  0.6496  2.4819

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  50.1251     0.2305 217.462  < 2e-16 ***
RNAi1        -2.4770     0.2627  -9.427 2.07e-12 ***
Drug1        -0.2253     0.2627  -0.857    0.396
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9282 on 47 degrees of freedom
Multiple R-squared:  0.6547,    Adjusted R-squared:   0.64
F-statistic: 44.55 on 2 and 47 DF,  p-value: 1.406e-11
```

```r
# Analyze data with Linear Regression model
Data2_lm1 <- lm(gene_exp ~ RNAi, data)
summary(Data2_lm1)

Data2_lm2 <- lm(gene_exp ~ RNAi + Drug, data)
summary(Data2_lm2)

Data2_lm3 <- lm(gene_exp ~ RNAi + Drug + RNAi*Drug, data)
summary(Data2_lm3)
```

```
> summary(Data2_lm3)

Call:
lm(formula = gene_exp ~ RNAi + Drug + RNAi * Drug, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-1.6885 -0.5726 -0.1925  0.6492  2.4823

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 50.124682   0.270848 185.066  < 2e-16 ***
RNAi1       -2.476247   0.375599  -6.593 3.73e-08 ***
Drug1       -0.224498   0.375599  -0.598    0.553
RNAi1:Drug1 -0.001577   0.531177  -0.003    0.998
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9382 on 46 degrees of freedom
Multiple R-squared:  0.6547,    Adjusted R-squared:  0.6322
F-statistic: 29.07 on 3 and 46 DF,  p-value: 1.076e-10
```

Comparing linear regression models with ANOVA

```r
#Find the best Linear Regression model
ANOVA_fit1 <- anova(Data2_lm1, Data2_lm2)
ANOVA_fit1

ANOVA_fit2 <- anova(Data2_lm1, Data2_lm3)
ANOVA_fit2

ANOVA_fit3 <- anova(Data2_lm2, Data2_lm3)
ANOVA_fit3

ANOVA_fit4 <- anova(Data2_lm1, Data2_lm2, Data2_lm3)
ANOVA_fit4
```

```
> ANOVA_fit1
Analysis of Variance Table

Model 1: gene_exp ~ RNAi
Model 2: gene_exp ~ RNAi + Drug
  Res.Df    RSS Df Sum of Sq F Pr(>F)
1     48 65.599
2     48 65.599  0         0
```

```
> ANOVA_fit2
Analysis of Variance Table

Model 1: gene_exp ~ RNAi
Model 2: gene_exp ~ RNAi + Drug + RNAi * Drug
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     48 65.599
2     47 65.251  1   0.34882 0.2513 0.6185
```

Comparing linear regression models with ANOVA

```r
#Find the best Linear Regression model
ANOVA_fit1 <- anova(Data2_lm1, Data2_lm2)
ANOVA_fit1

ANOVA_fit2 <- anova(Data2_lm1, Data2_lm3)
ANOVA_fit2

ANOVA_fit3 <- anova(Data2_lm2, Data2_lm3)
ANOVA_fit3

ANOVA_fit4 <- anova(Data2_lm1, Data2_lm2, Data2_lm3)
ANOVA_fit4
```

```
> ANOVA_fit3
Analysis of Variance Table

Model 1: gene_exp ~ RNAi + Drug
Model 2: gene_exp ~ RNAi + Drug + RNAi * Drug
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     48 65.599
2     47 65.251  1   0.34882 0.2513 0.6185
```

```
> ANOVA_fit4
Analysis of Variance Table

Model 1: gene_exp ~ RNAi
Model 2: gene_exp ~ RNAi + Drug
Model 3: gene_exp ~ RNAi + Drug + RNAi * Drug
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     48 65.599
2     48 65.599  0   0.00000
3     47 65.251  1   0.34882 0.2513 0.6185
```

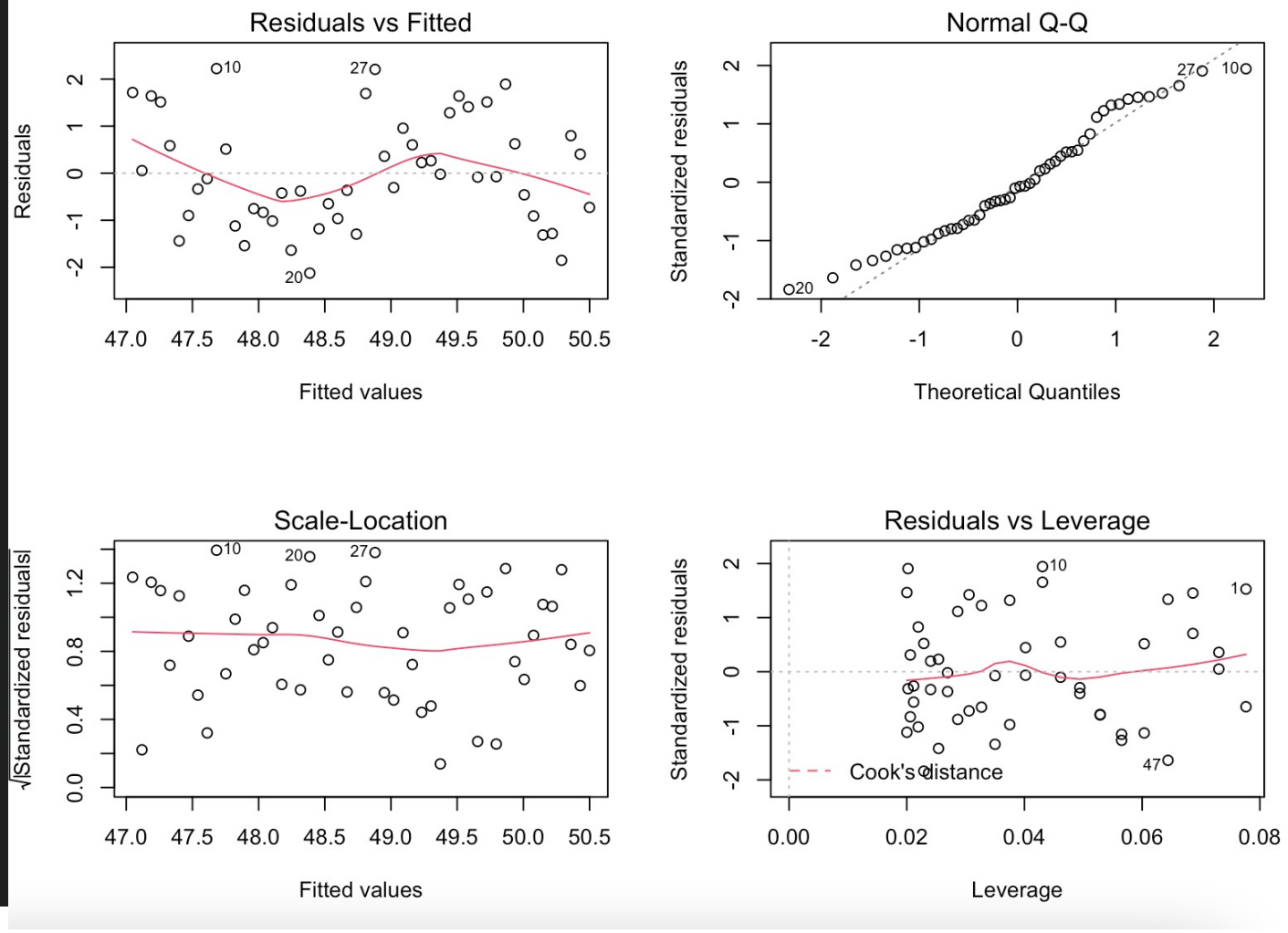
Comparing linear regression models with ANOVA

```r
model.set <- list(Data2_lm1, Data2_lm2, Data2_lm3)
model.names <- c("lm1", "lm2", "lm3")
aictab(model.set, modnames = model.names)
```

```
Model selection based on AICc:

    K   AICc Delta_AICc AICcWt Cum.Wt     LL
lm1 3 161.99        0.0   0.43    0.43 -77.74
lm2 3 161.99        0.0   0.43    0.85 -77.74
lm3 4 164.09        2.1   0.15    1.00 -77.60
```

Check for homoscedasticity

```
# Check for homoscedasticity
par(mfrow=c(2,2))
plot(Data2_lm1)
par(mfrow=c(1,1))

par(mfrow=c(2,2))
plot(Data2_lm2)
par(mfrow=c(1,1))

par(mfrow=c(2,2))
plot(Data2_lm3)
par(mfrow=c(1,1))
```



To summarize:

Both ANOVA and Linear Regression are linear models.

In ANOVA, the independent variable(s) is qualitative, and the dependent variable is quantitative.

In Linear Regression, both the independent and dependent variables are quantitative.

ANOVA is used to test if group means are different while Linear Regression is used to determine the correlation between variables

Questions?