General linear models

	Linear Model	Common name
✓	$Y = \mu + X$	Simple linear regression
✓	$Y = \mu + A_1$	One-factorial (one-way) ANOVA
	$Y = \mu + A_1 + A_2 + A_1 \times A_2$	Two-factorial (two-way) ANOVA
	$Y = \mu + A_1 + X + (A_1 \times X)$	Analysis of Covariance (ANCOVA)
	$Y = \mu + X_1 + X_2 + \ldots + X_p$	Multiple regression
	$Y = \mu + A_1 + g + A_1 \times g$	Mixed model ANOVA
	$(Y_1, Y_{2,\ldots} Y_r)$ $= \mu + X_1 + X_2 + \ldots + X_p$	Multi-response models

Y (response) is a continuous variable
X (predictor) is a continuous variable
A represents categorical predictors (factors)
g represents groups of data (more on this later)

Some types of (Analysis of Variance – ANOVA) designs:

Single-factor ANOVA (Intro stats)

Factorial designs (crossed) – today

Mixed models

Research question (my own fictional example; real examples will be seen in the next lecture and tutorials):

Do exercise and diet affect weight loss?

How would you set a study to test this question?

Why fictional? The context of the problem itself seems to be easier to understand than more “biological” applications!

Study: Individuals are followed for one month to assess weight loss with and without exercise and diet.

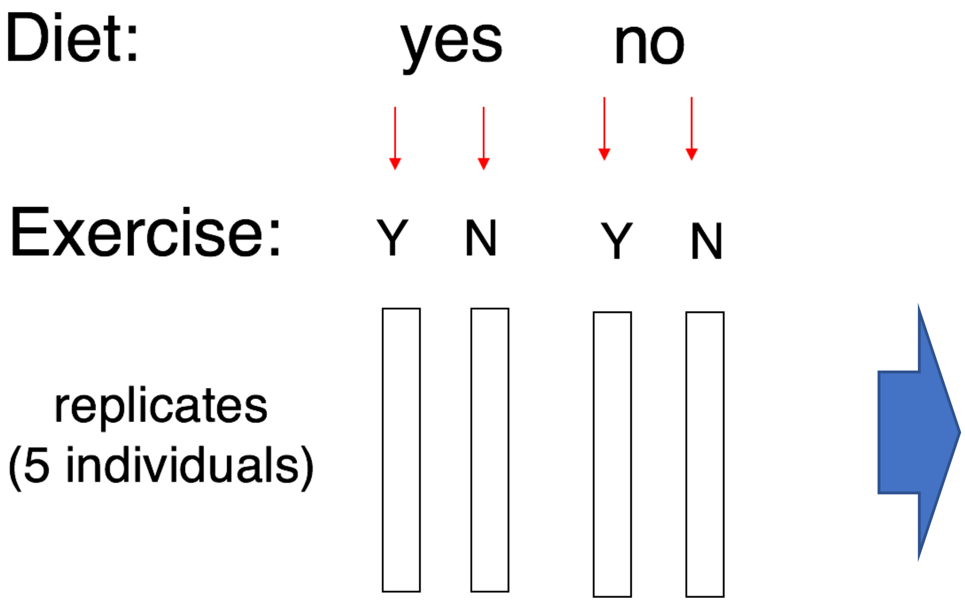
Factorial ANOVA - always involves one continuous variable (i.e., response variable = weight loss) and two or more categorical (factors) variables (exercise and diet).

Factors: exercise and diet (two-factorial).

In this example, **exercise** and **diet** are factors with two levels or groups (Yes/No).

Response variable: weight loss.

Data structure in a csv file

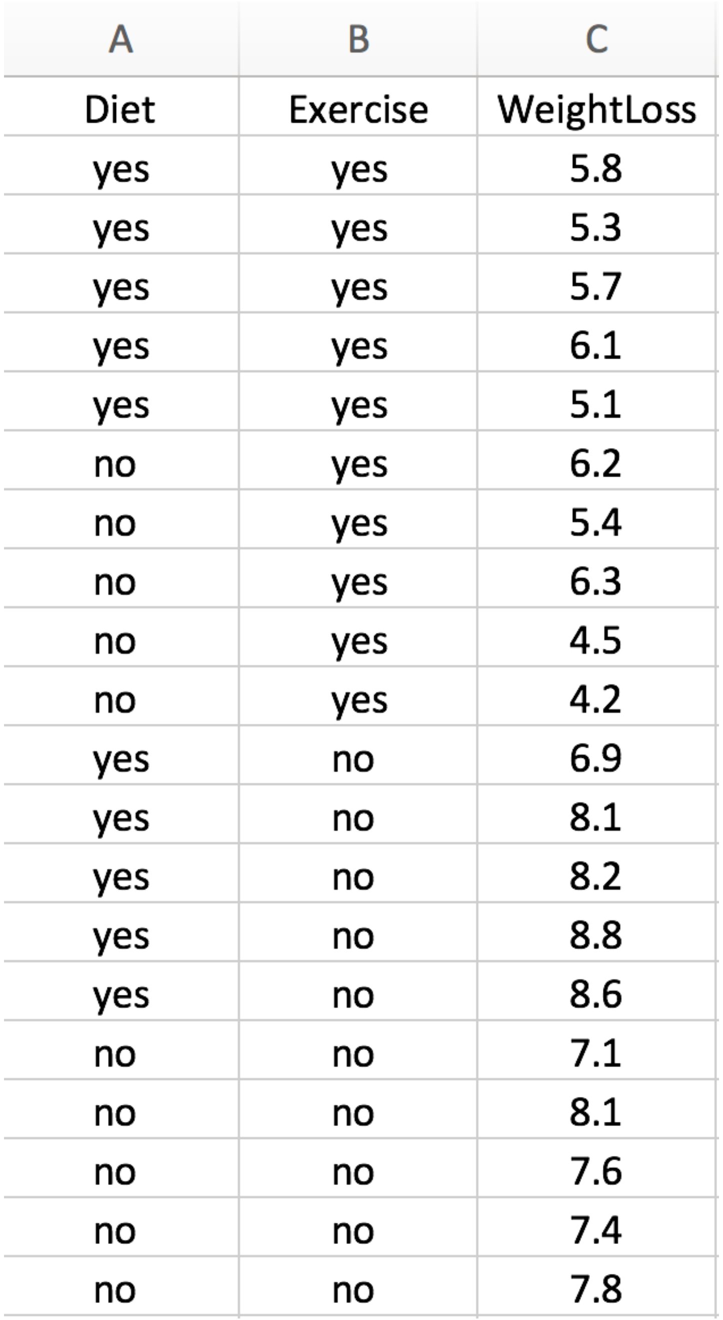


Diet: yes no

Exercise: Y N Y N

replicates
(5 individuals)

Weight loss: start weight - end weight (in pounds)



A	B	C
Diet	Exercise	WeightLoss
yes	yes	5.8
yes	yes	5.3
yes	yes	5.7
yes	yes	6.1
yes	yes	5.1
no	yes	6.2
no	yes	5.4
no	yes	6.3
no	yes	4.5
no	yes	4.2
yes	no	6.9
yes	no	8.1
yes	no	8.2
yes	no	8.8
yes	no	8.6
no	no	7.1
no	no	8.1
no	no	7.6
no	no	7.4
no	no	7.8

Do exercise and diet affect weight loss?

Let’s elaborate on this question further:

Main effects

- Are the differences in weight loss only due to DIET alone?
- Are the differences in weight loss only due to EXERCISE alone?

Interaction

- Does the effect of DIET on weight loss depend on EXERCISE? In other words, are the differences in weight loss attributable to some combinations of exercise and diet? (e.g., the biggest weight loss compared to any other combination of diet and exercise was observed when individuals both dieted and exercised).

Treatments

Main effects:

Diet - two treatments (yes/no).

Exercise - two treatments (yes/no).

Possible sources of statistical interactions:

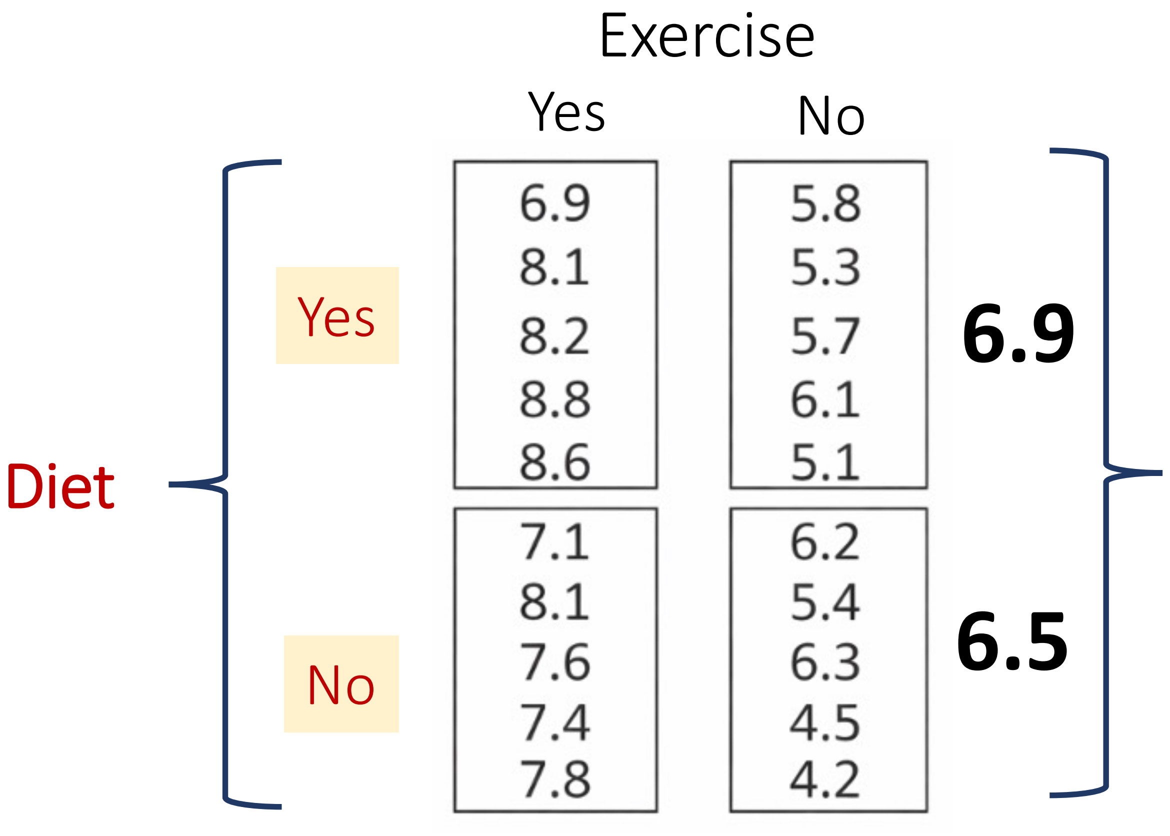
Combination of diet and exercise treatments - four pairwise combinations of means:

1) No exercise but diet.
2) Exercise but no diet.
3) No exercise and no diet.
4) Exercise and diet.

Does DIET alone (main effect) influence weight loss?

Statistically, is the difference between 6.9 and 6.5 larger than expected from random sampling variation alone?

In other words, we test whether the observed difference between sample means exceeds what would be expected if both DIET regimes (yes/no) came from populations with the same underlying mean (the null hypothesis). If the difference exceeds (e.g., P < alpha level), then we generated evidence that diet alone influences weight loss.



		Exercise		
		Yes	No	
Diet	Yes	6.9 8.1 8.2 8.8 8.6	5.8 5.3 5.7 6.1 5.1	**6.9**
	No	7.1 8.1 7.6 7.4 7.8	6.2 5.4 6.3 4.5 4.2	**6.5**

Row means
(Diet independent of exercise)

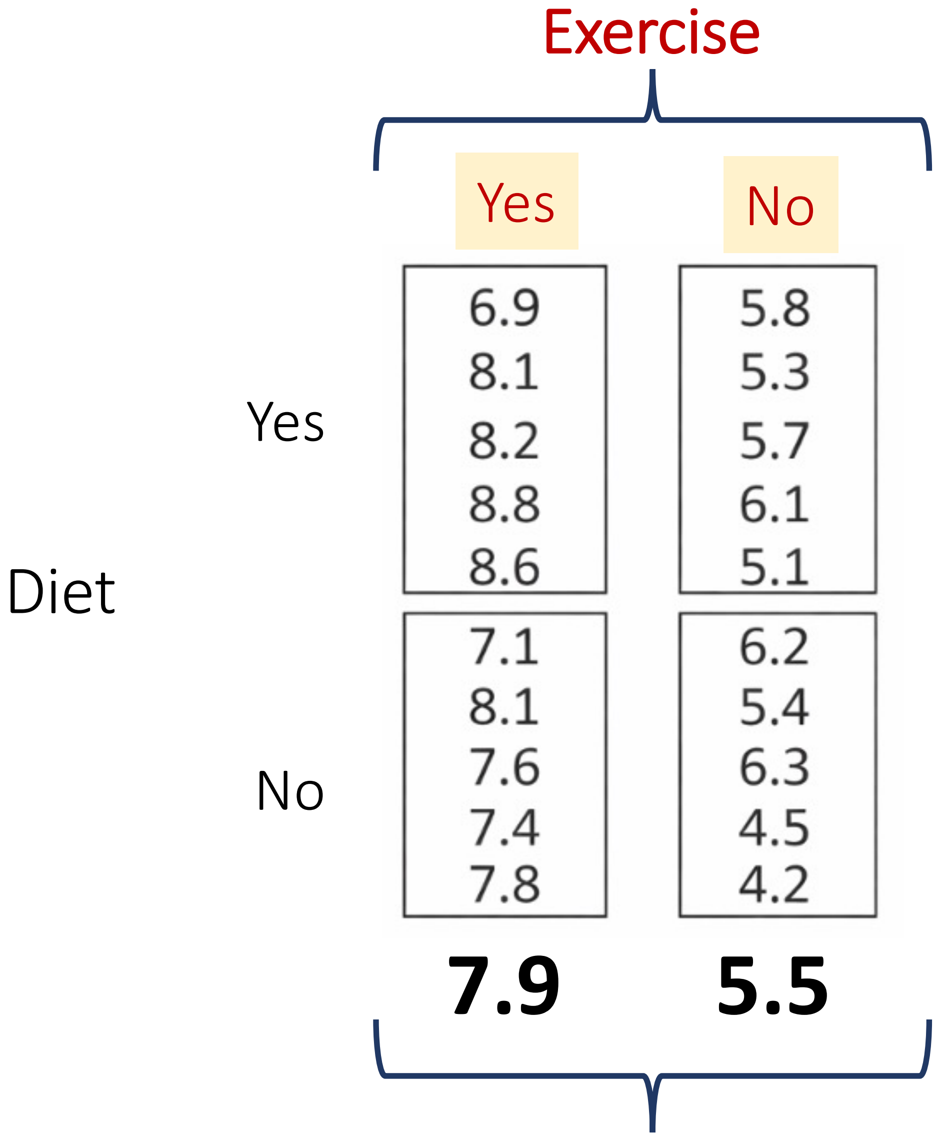
Marginal effect of Diet
(6.9 kg – 6.5 kg)

The **marginal effect of Diet** is the difference between the mean response for *Diet = Yes* and *Diet = No*, **averaged over exercise**.

Does DIET alone (main effect) influence weight loss?

Statistically, is the difference between 7.9 and 5.5 larger than expected from random sampling variation alone?

In other words, we test whether the observed difference between sample means exceeds what would be expected if both EXERCISE regimes (yes/no) came from populations with the same underlying mean (the null hypothesis). If the difference exceeds (e.g., P < alpha level), then we generated evidence that exercise alone influence weight loss.



		Exercise	
		Yes	No
Diet	Yes	6.9 8.1 8.2 8.8 8.6	5.8 5.3 5.7 6.1 5.1
	No	7.1 8.1 7.6 7.4 7.8	6.2 5.4 6.3 4.5 4.2
		7.9	**5.5**

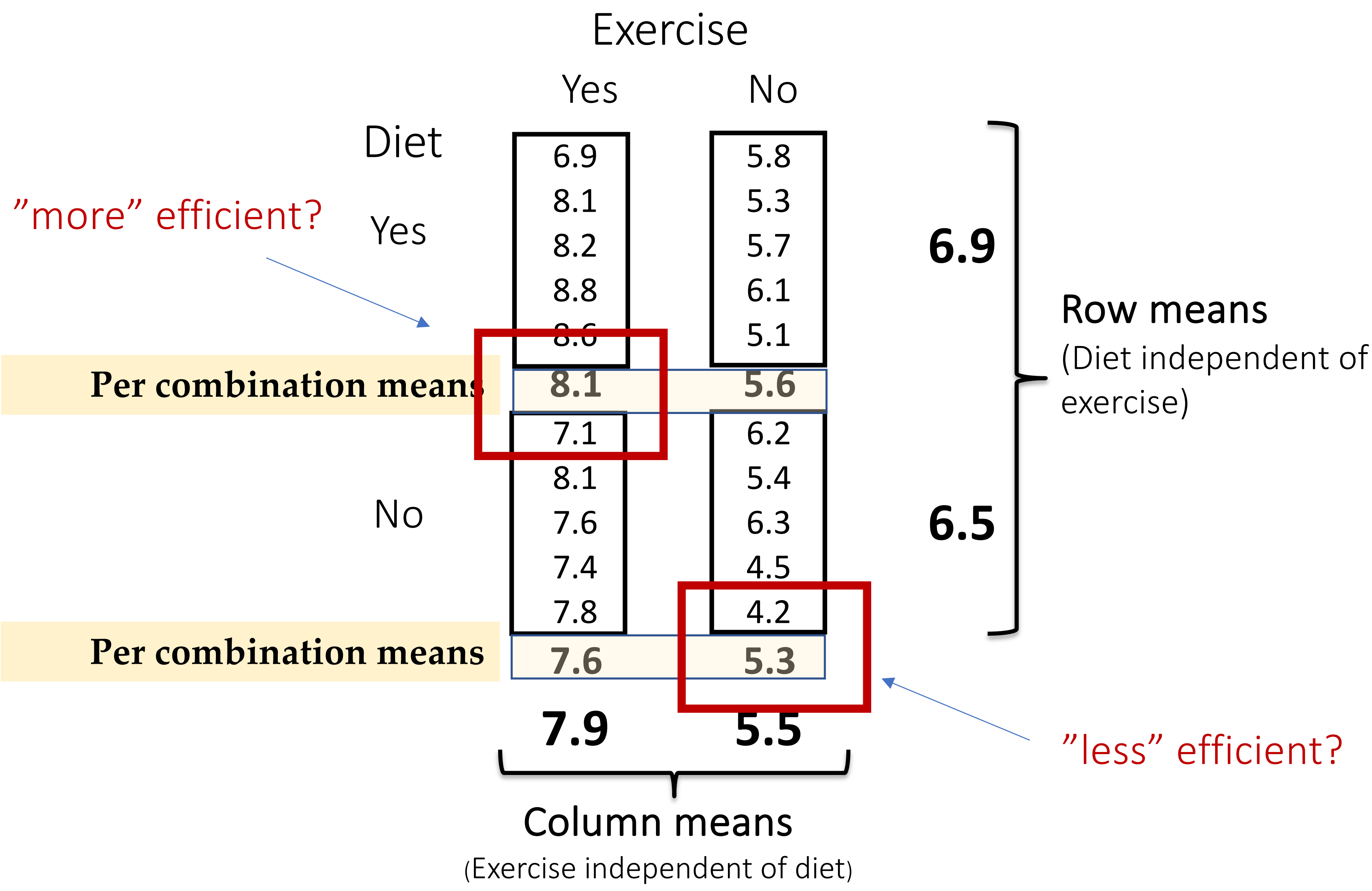
Columns means

(Exercise independent of diet)

Marginal effect of Exercise
(7.9 kg – 5.5 kg)

The **marginal effect of Exercise** is the difference between the mean response for *Exercise = Yes* and *Diet = No*, **averaged over diet**.

Does the effect of diet depend on exercise? (or vice versa; i.e., is there an interaction between exercise and diet that affects weight loss?) - Does the marginal effect of exercise differ (depend) among diets?



		Exercise		
Diet		Yes	No	
Yes		6.9	5.8	
		8.1	5.3	
		8.2	5.7	6.9
		8.8	6.1	
		8.6	5.1	
	Per combination means	8.1	5.6	
No		7.1	6.2	
		8.1	5.4	
		7.6	6.3	6.5
		7.4	4.5	
		7.8	4.2	
	Per combination means	7.6	5.3	
		7.9	**5.5**	

Row means

(Diet independent of exercise)

Column means

(Exercise independent of diet)

Stating the 3 possible sets of statistical hypotheses in a two-factorial design:

Does *dieting* affect weight loss? DIET (main effect 1)

H_0: There is no difference between diet treatments in mean weight loss.

H_A: There is a difference between diet treatments in mean weight loss.

Stating the 3 possible sets of statistical hypotheses in a two-factorial design:

Does *exercising* affect weight loss? EXERCISE (main effect 2)

H_0: There is no difference between exercise treatments in mean weight loss.

H_A: There is a difference between exercise treatments in mean weight loss.

Stating the 3 possible sets of statistical hypotheses in a two-factorial design:

Are the differences in weight loss attributable to some combinations of exercise and diet? (interaction effect)

H_0: The effect of diet on weight loss does not depend on exercise (*or vice versa*).

H_A: The effect of diet on weight loss depends on exercise (*or vice versa*).

Type of effects in this study:

Fixed: The levels in a factor were specifically chosen by the researcher (diet – yes/not and exercise – yes/no)

Note: The typical ANOVA design (simple or factorial) is conducted assuming a fixed design (we will see other designs later in the course).

ANOVA Table

Source of variation	Df	SS	Mean SS	F value	Prob
Diet	1	0.800	0.800	1.8089	0.1974
Exercise	1	28.800	28.800	65.1215	<0.0000001
Diet x Exercise	1	0.072	0.072	0.1628	0.6919
residuals	16	7.076	0.442		

H_0: There is no difference between diet treatments in mean weight loss.
H_A: There is a difference between diet treatments in mean weight loss.

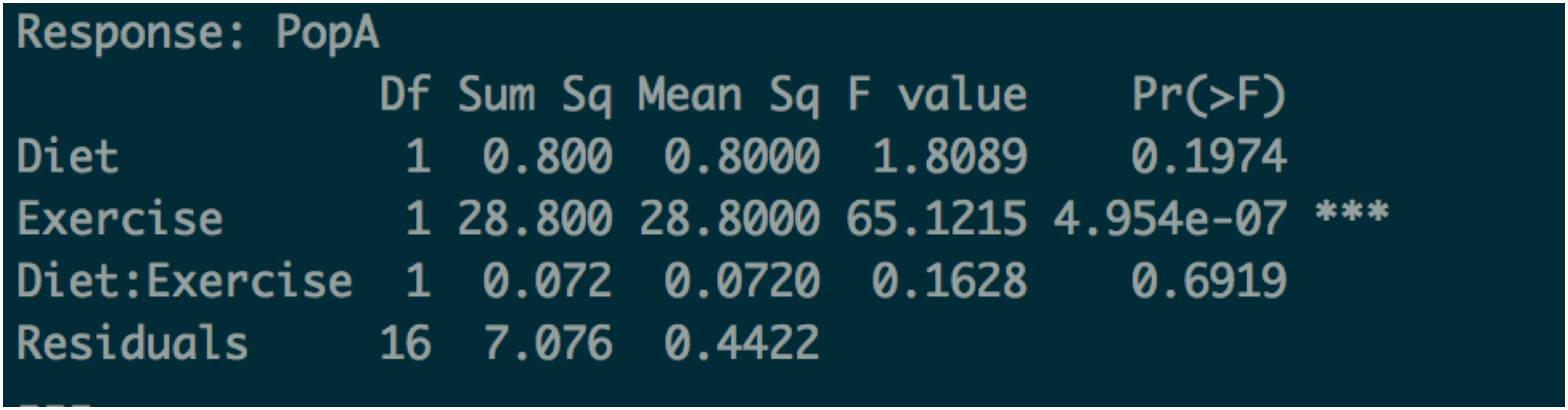
H_0: There is no difference between exercise treatments in mean weight loss.
H_A: **There is a difference between exercise treatments in mean weight loss.**

H_0: The effect of diet on weight loss does not depend on exercise (*or vice versa*).
H_A: The effect of diet on weight loss depends on exercise (*or vice versa*).

Research conclusion: *Weight loss differs strongly between exercise treatments, but not between diets, and there is no evidence that diet and exercise interact.*

In other words, exercise has a significant main effect on weight loss, while diet does not, and the effect of exercise is consistent across diets.

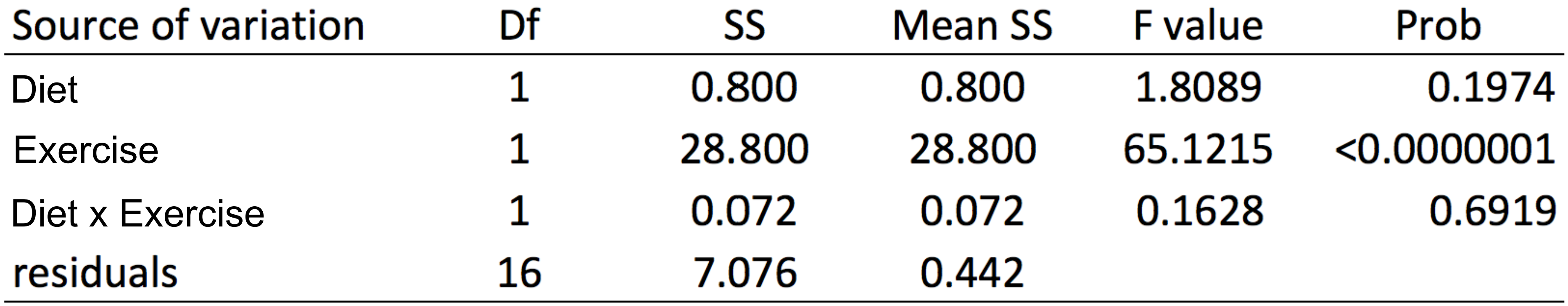
ANOVA Table (R) versus publication quality



```
Response: PopA
               Df Sum Sq Mean Sq F value    Pr(>F)    
Diet            1  0.800  0.8000  1.8089    0.1974    
Exercise        1 28.800 28.8000 65.1215 4.954e-07 ***
Diet:Exercise  1  0.072  0.0720  0.1628    0.6919    
Residuals      16  7.076  0.4422                      
```

Source of variation	Df	SS	Mean SS	F value	Prob
Diet	1	0.800	0.800	1.8089	0.1974
Exercise	1	28.800	28.800	65.1215	<0.0000001
Diet x Exercise	1	0.072	0.072	0.1628	0.6919
residuals	16	7.076	0.442		

ANOVA Table (details on degrees of freedom)



Source of variation	Df	SS	Mean SS	F value	Prob
Diet	1	0.800	0.800	1.8089	0.1974
Exercise	1	28.800	28.800	65.1215	<0.0000001
Diet x Exercise	1	0.072	0.072	0.1628	0.6919
residuals	16	7.076	0.442		

df (diet) = number of levels (k) - 1 = 2 - 1 = 1

df (exercise) = number of levels (m) - 1 = 2 - 1 = 1

df (Interaction) = (m - 1).(k - 1) = (2 - 1).(2 - 1) = 1

df (residuals) = (N - m - k) = (20 - 2 - 2) = 16

N = total number of observations across all factors and levels

Next:

1) Real examples of two-way ANOVA designs.

2) Plotting and understanding significant interaction terms.

3) How to test for assumptions (one-way and multi-factorial ANOVA).

4) Identifying which pairs of means significantly differ to find the meaningful interactions (e.g., mean of weight loss with no exercise *versus* mean of weight loss with diet).

ANOVA Table

Source of variation	Df	SS	Mean SS	F value	Prob
Diet	1	0.800	0.800	1.8089	0.1974
Exercise	1	28.800	28.800	65.1215	<0.0000001
Diet x Exercise	1	0.072	0.072	0.1628	0.6919
residuals	16	7.076	0.442		

H_0: There is no difference between diet treatments in mean weight loss.
H_A: There is a difference between diet treatments in mean weight loss.

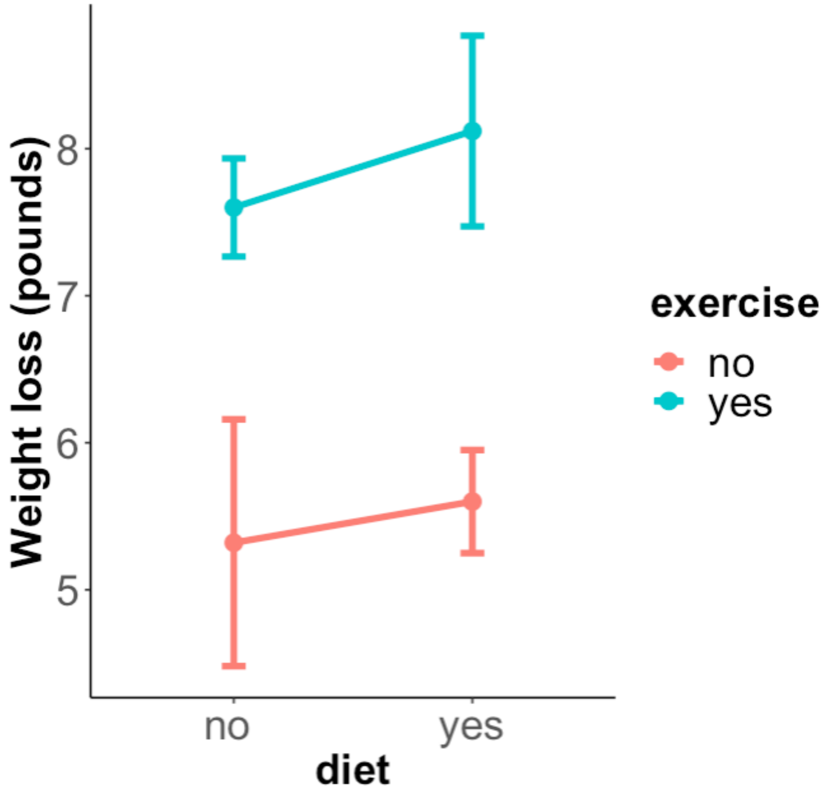
H_0: There is no difference between exercise treatments in mean weight loss.
H_A: **There is a difference between exercise treatments in mean weight loss.**

H_0: The effect of diet on weight loss does not depend on exercise (*or vice versa*).
H_A: The effect of diet on weight loss depends on exercise (*or vice versa*).

Research conclusion: *Weight loss differs strongly between exercise treatments, but not between diets, and there is no evidence that diet and exercise interact.*

Only exercise affects weight loss!
BUT HOW? Exercise increases weight loss ($P<0.0000001$)

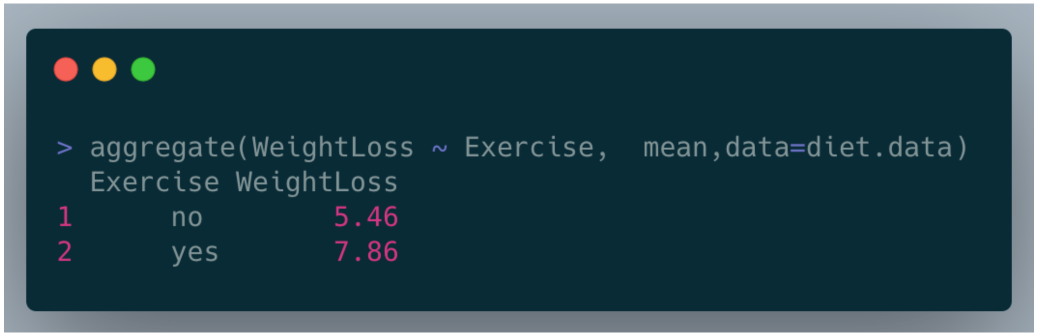
Interaction plot & 95% confidence intervals



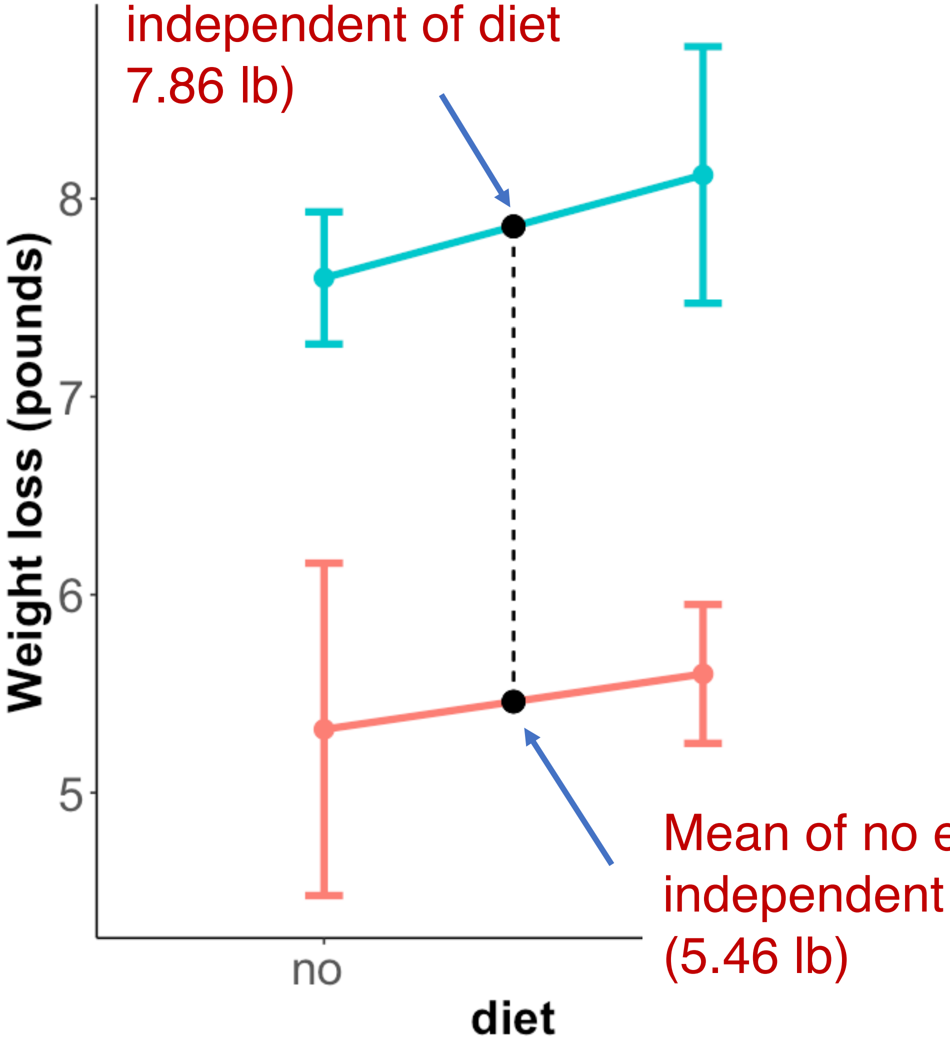
Interaction plots provide the most intuitive way to understand results from complex factorial studies. Statistical significance tests will be introduced later to formally assess whether the visual patterns we observe are meaningful.

Only exercise affects weight loss!
BUT HOW? Exercise increases weight loss (P<0.0000001)

Understanding interaction plots



```
> aggregate(WeightLoss ~ Exercise,  mean,data=diet.data)
  Exercise WeightLoss
1       no       5.46
2      yes       7.86
```



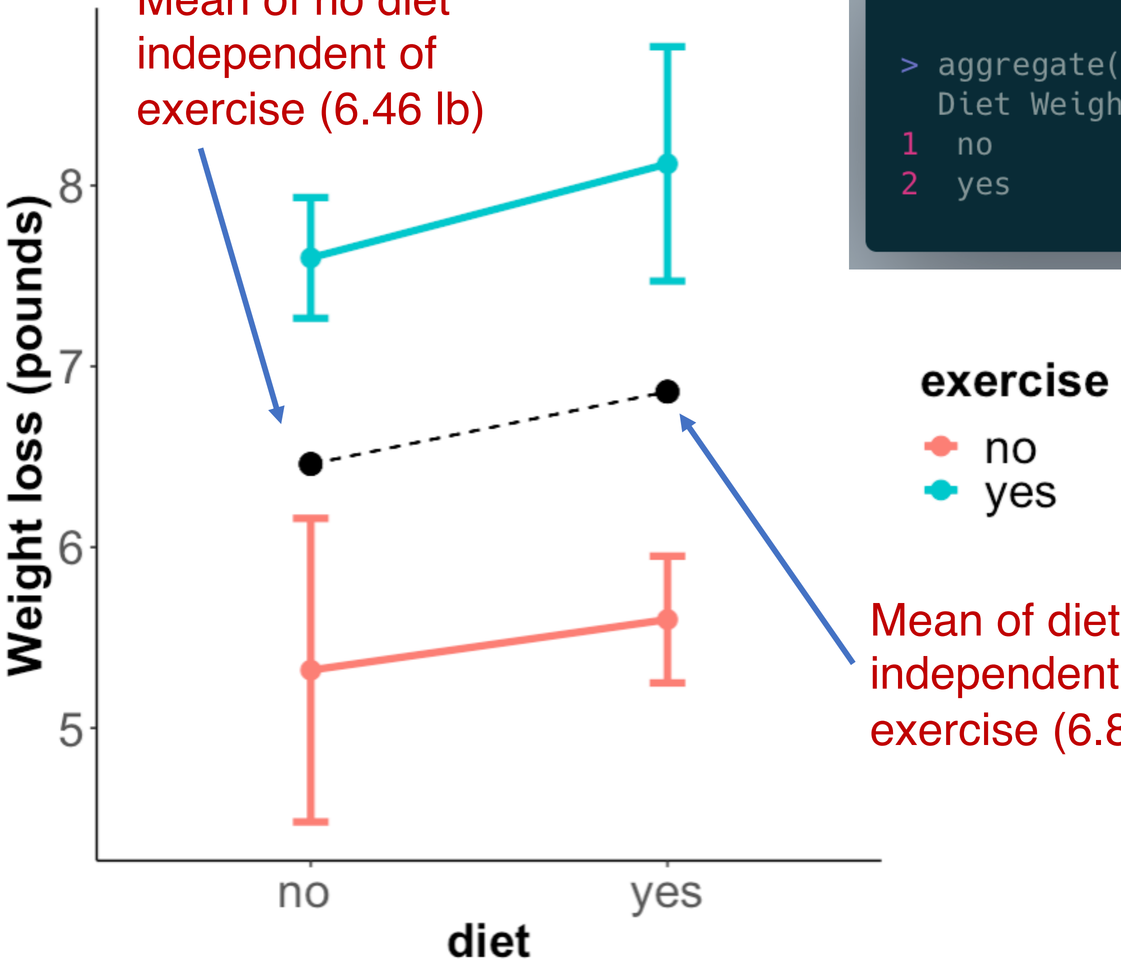
Mean of exercise independent of diet 7.86 lb)

Mean of no exercise independent of diet (5.46 lb)

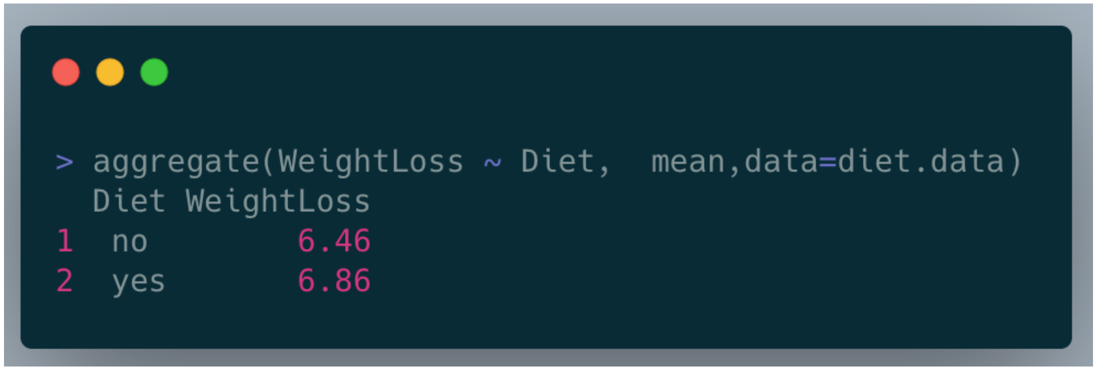
Note: Dots on the line and dashed lines are included for clarity only; standard interaction plots use solid lines.

Diet did not affect weight loss! (P=0.1974), i.e., variation in mean (likely) due to sampling variation

Mean of no diet independent of exercise (6.46 lb)



Mean of diet independent of exercise (6.86 lb)



```
> aggregate(WeightLoss ~ Diet,  mean,data=diet.data)
  Diet WeightLoss
1  no        6.46
2  yes       6.86
```

There are five different possible outcomes from a two-way factorial ANOVA:

CASE 1: Only one main effect is significant (either DIET or EXERCISE). ✓

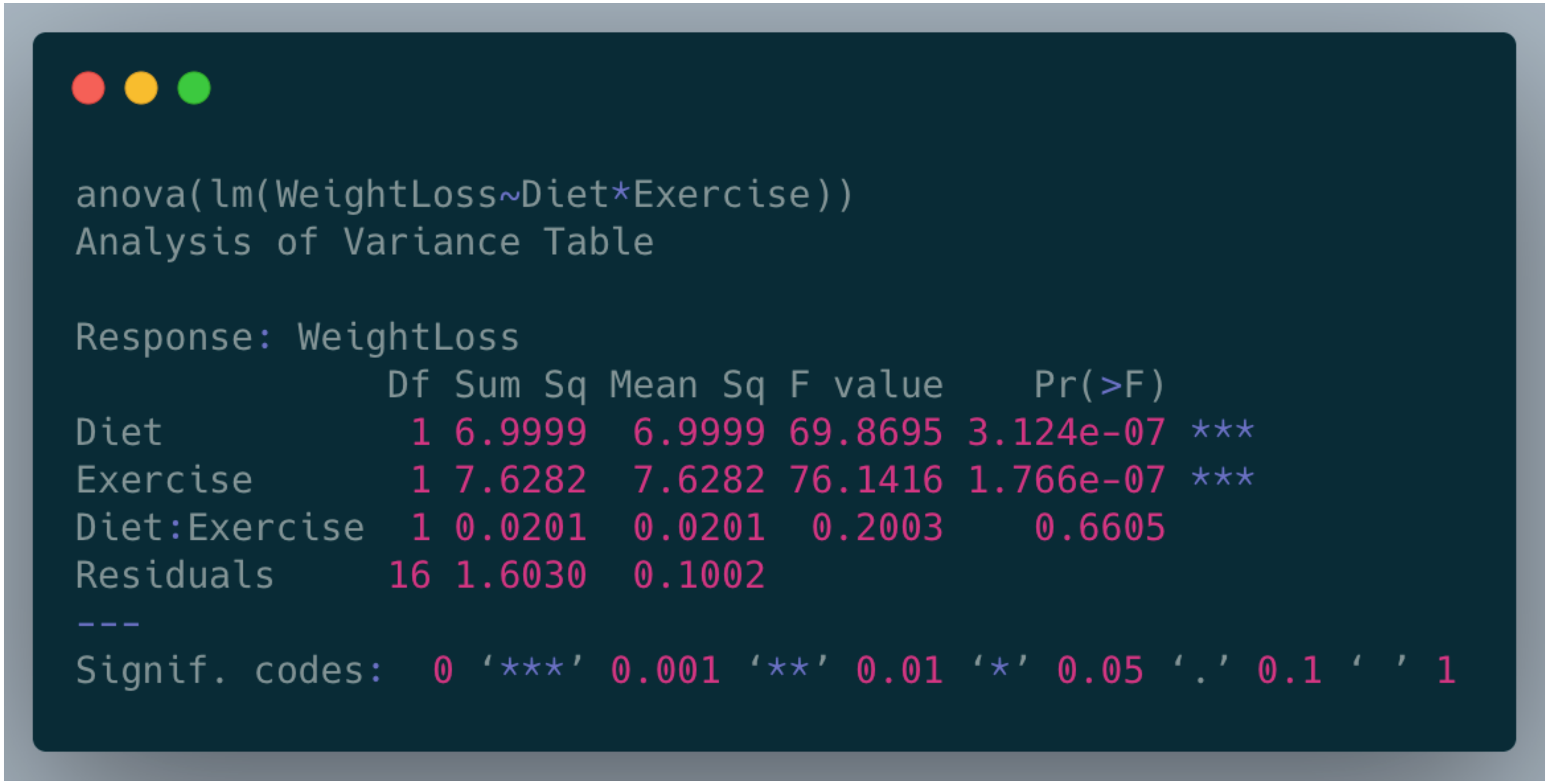
CASE 2: The two main effects are significant (both DIET AND EXERCISE) but not the interaction.

CASE 3: Only the interaction is significant.

CASE 4: One or both main factors are significant and the interaction as well.

CASE 5: No factor or interaction are significant (no need to cover this one; at least not graphically).

CASE 2: the two main effects are significant (DIET AND EXERCISE) but not the interaction.

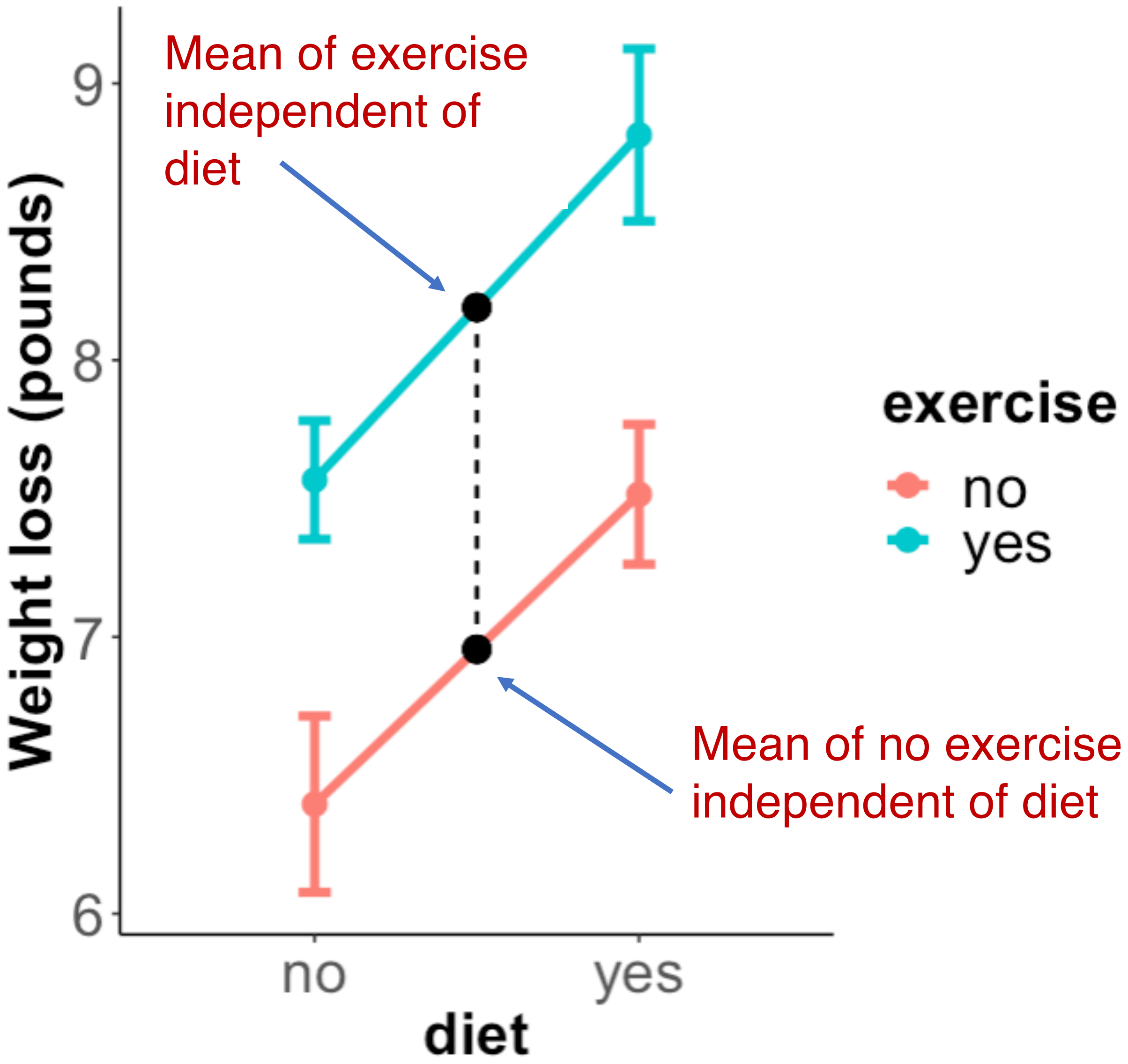


```
anova(lm(WeightLoss~Diet*Exercise))
Analysis of Variance Table

Response: WeightLoss
              Df Sum Sq Mean Sq F value    Pr(>F)    
Diet           1 6.9999  6.9999 69.8695 3.124e-07 ***
Exercise       1 7.6282  7.6282 76.1416 1.766e-07 ***
Diet:Exercise  1 0.0201  0.0201  0.2003    0.6605    
Residuals     16 1.6030  0.1002                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Note that I kept the “fictional study”, but I’ve created data for the different outcomes (cases).

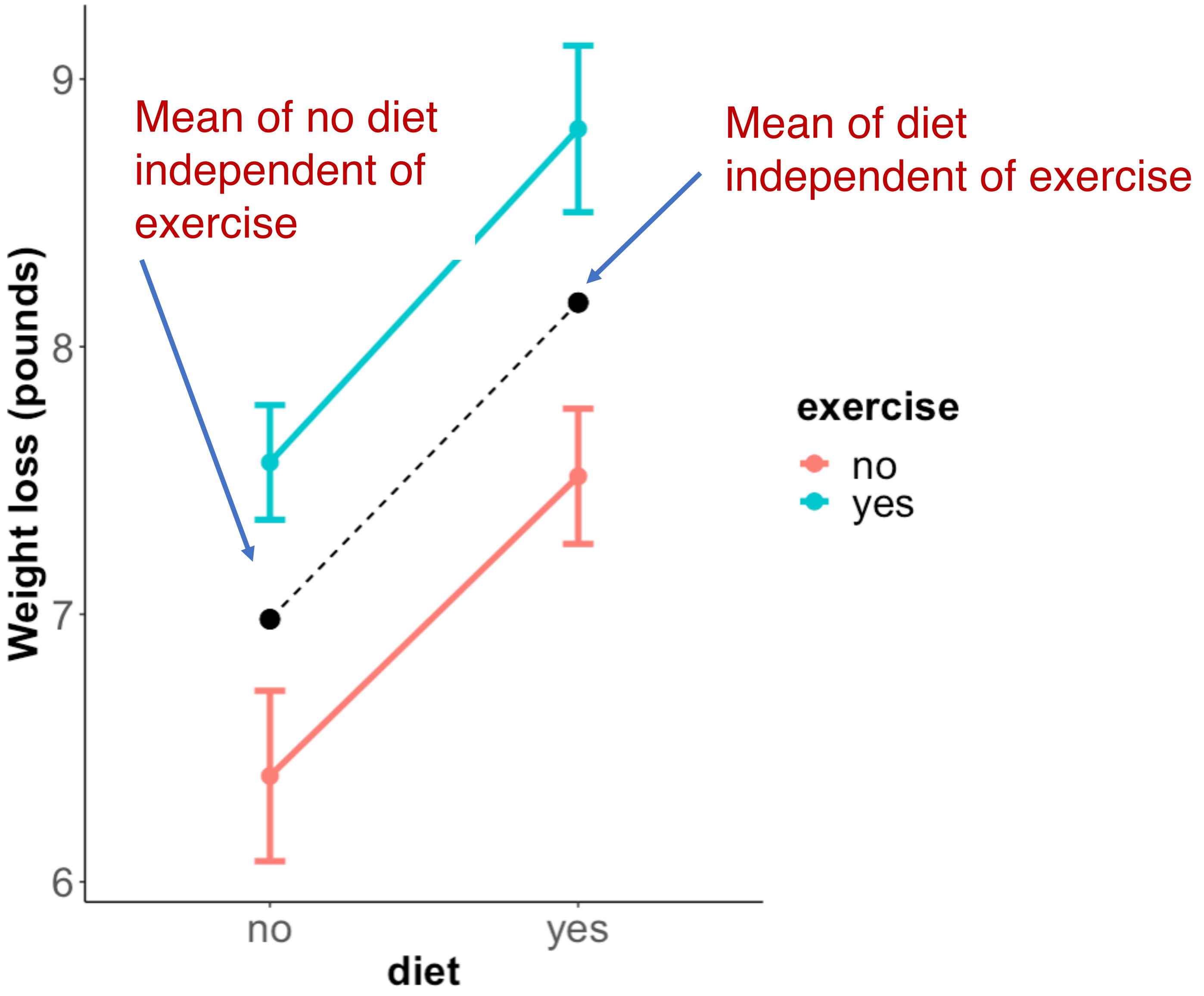
CASE 2: The two main effects are significant (DIET AND EXERCISE) but not the interaction.



Mean of exercise independent of diet

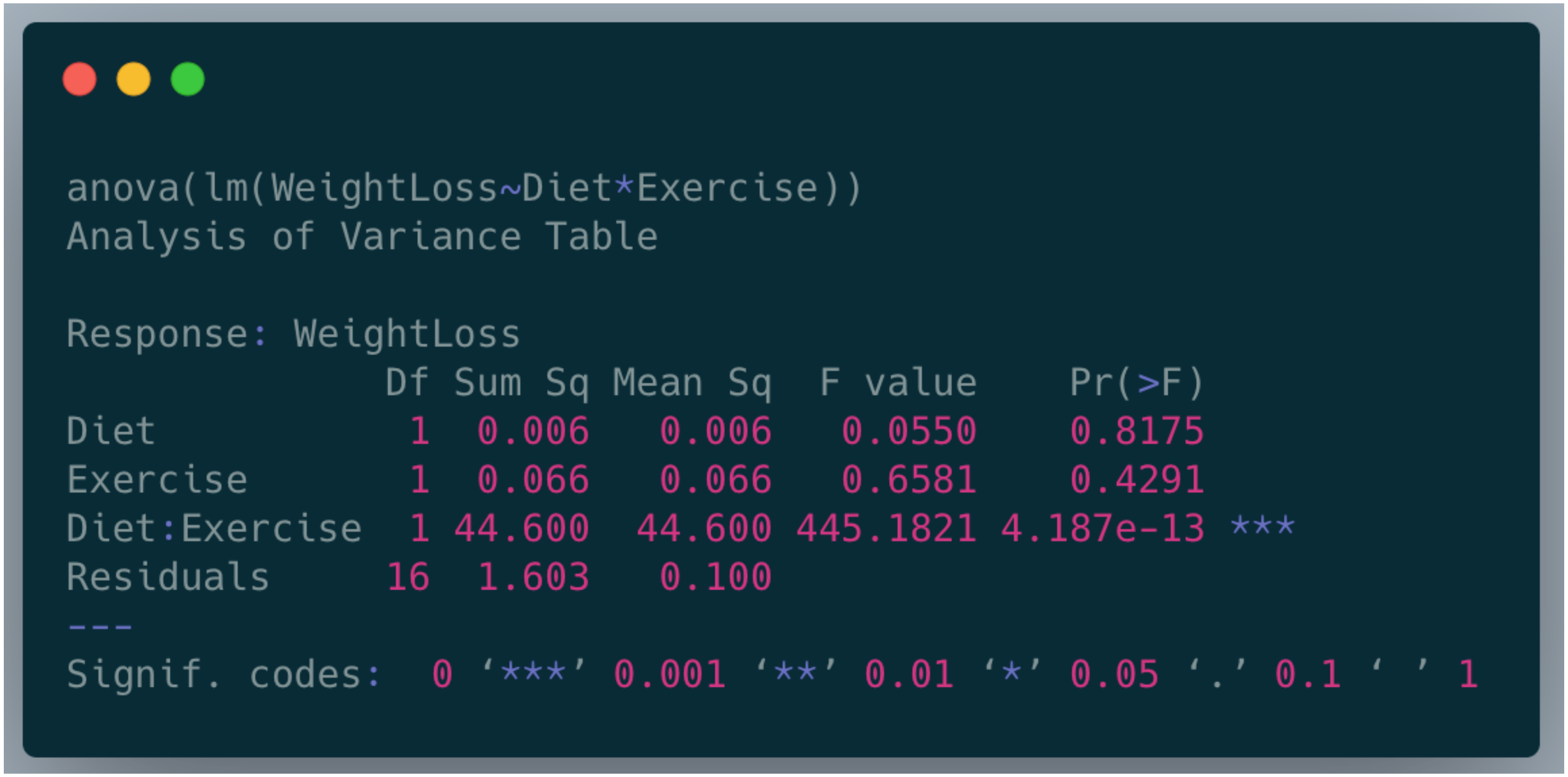
Mean of no exercise independent of diet

CASE 2: The two main effects are significant (DIET AND EXERCISE) but not the interaction.


Mean of no diet independent of exercise

Mean of diet independent of exercise

CASE 3: Only the interaction is significant

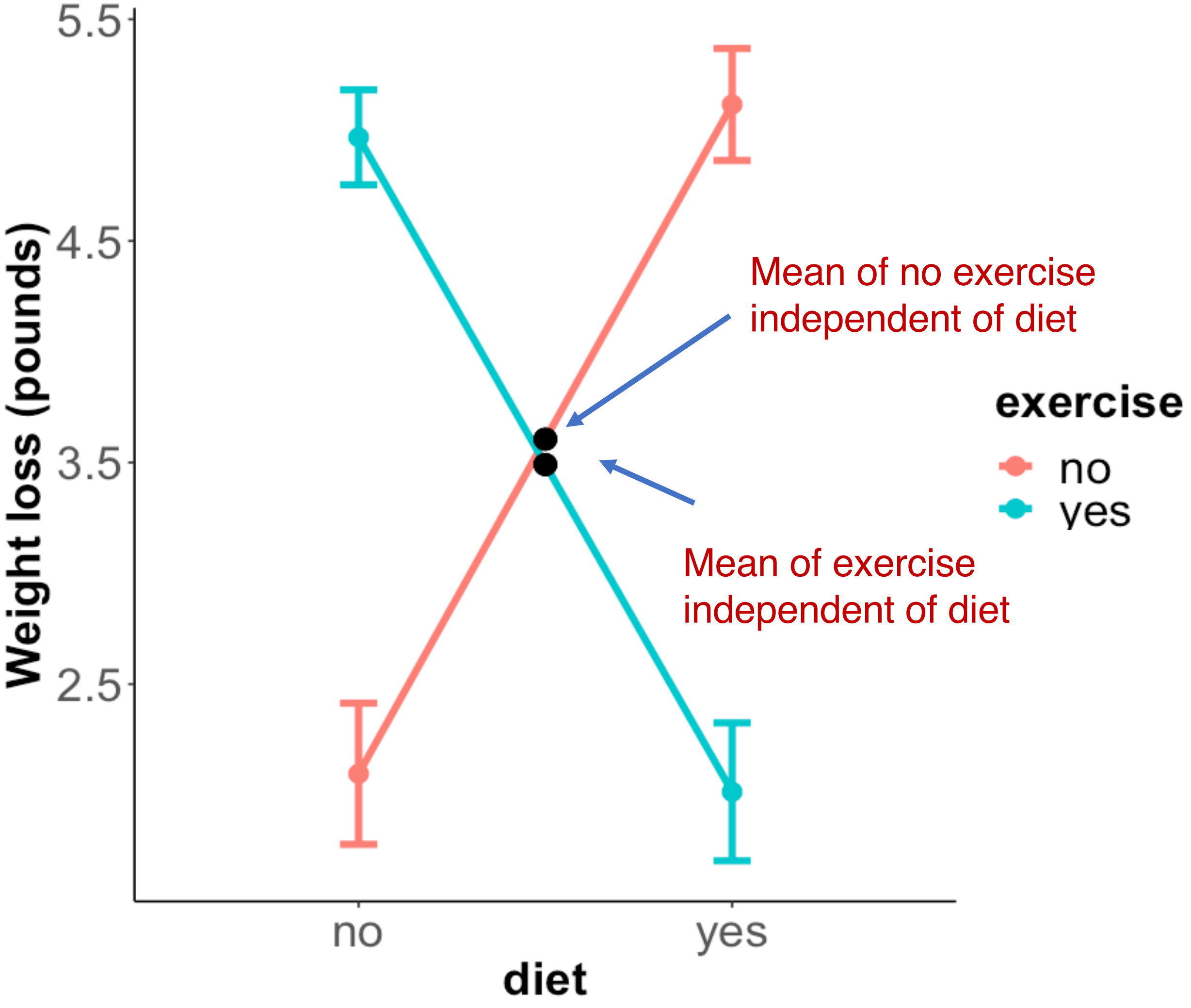


```
anova(lm(WeightLoss~Diet*Exercise))
Analysis of Variance Table

Response: WeightLoss
              Df Sum Sq Mean Sq  F value    Pr(>F)    
Diet           1  0.006   0.006   0.0550    0.8175    
Exercise       1  0.066   0.066   0.6581    0.4291    
Diet:Exercise  1 44.600  44.600 445.1821 4.187e-13 ***
Residuals     16  1.603   0.100                       
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Note that I kept the “fictional study”, but I’ve created different data set for the different possible outcomes (cases).

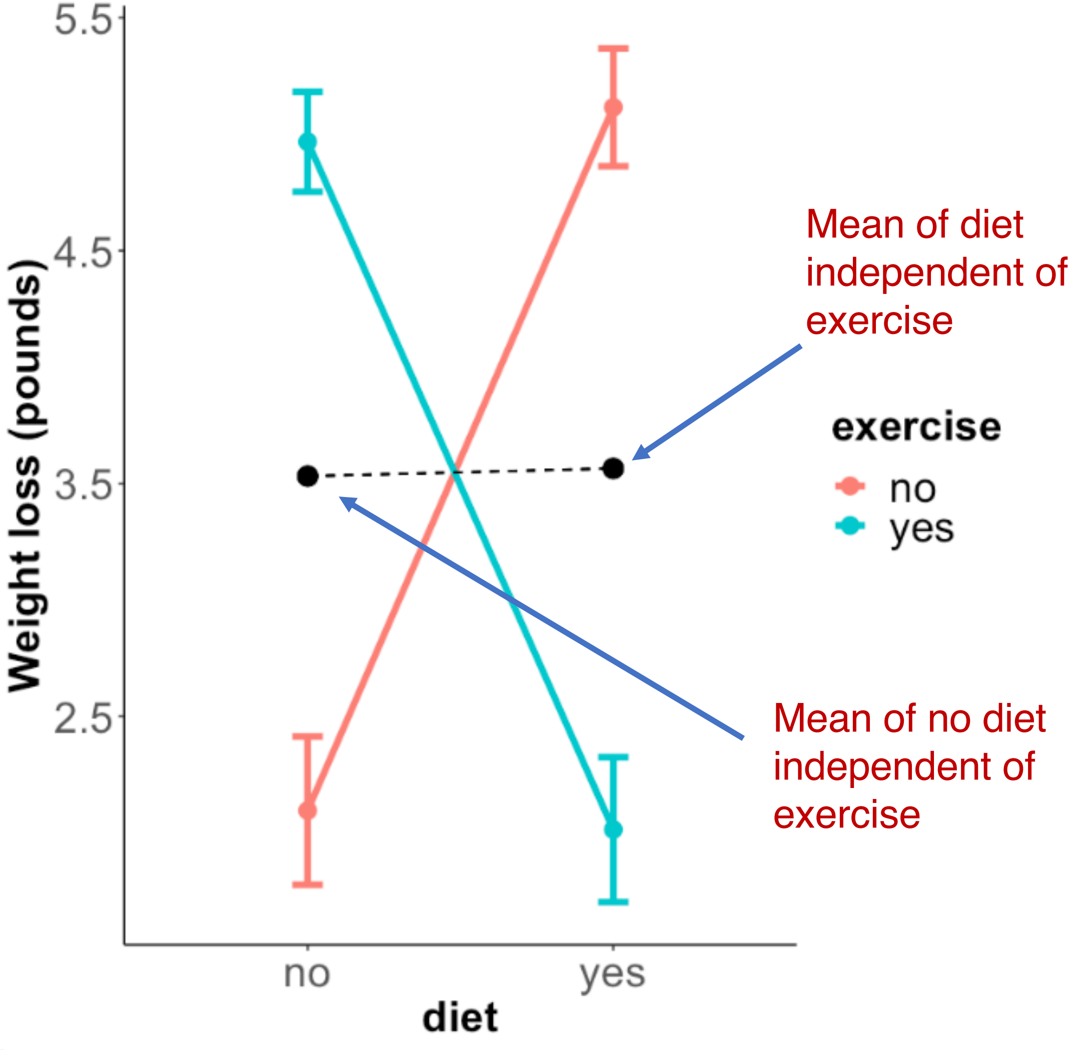
CASE 3: Only the interaction is significant, i.e., weight loss depends on the combinations of the levels of the main effects; greater when no diet and exercise OR when diet and no exercise.



Mean of no exercise independent of diet

Mean of exercise independent of diet

CASE 3: Only the interaction is significant, i.e., weight loss depends on the combinations of the levels of the main effects; greater when no diet and exercise OR when diet and no exercise.

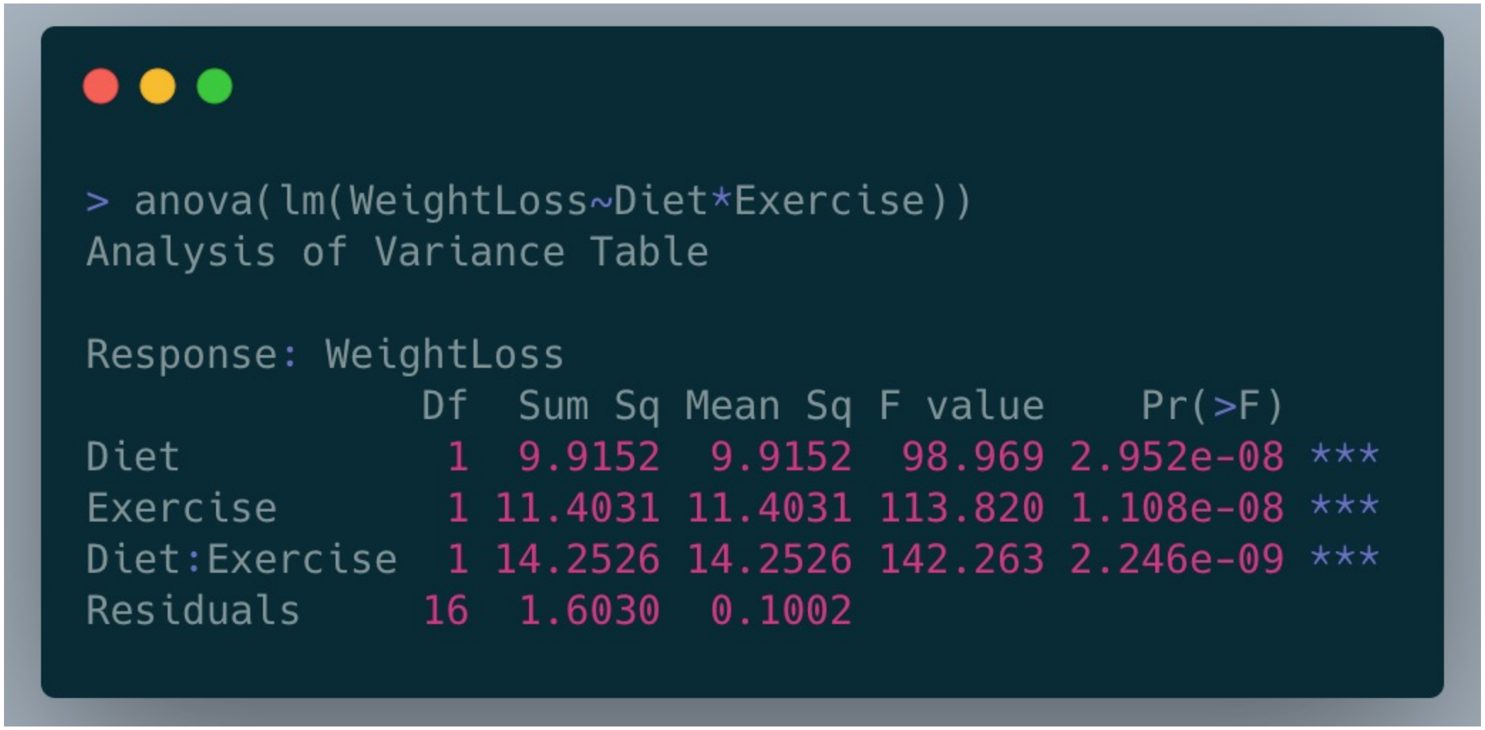


Mean of diet independent of exercise

Mean of no diet independent of exercise

CASE 4: One or both main factors are significant and the interaction as well.

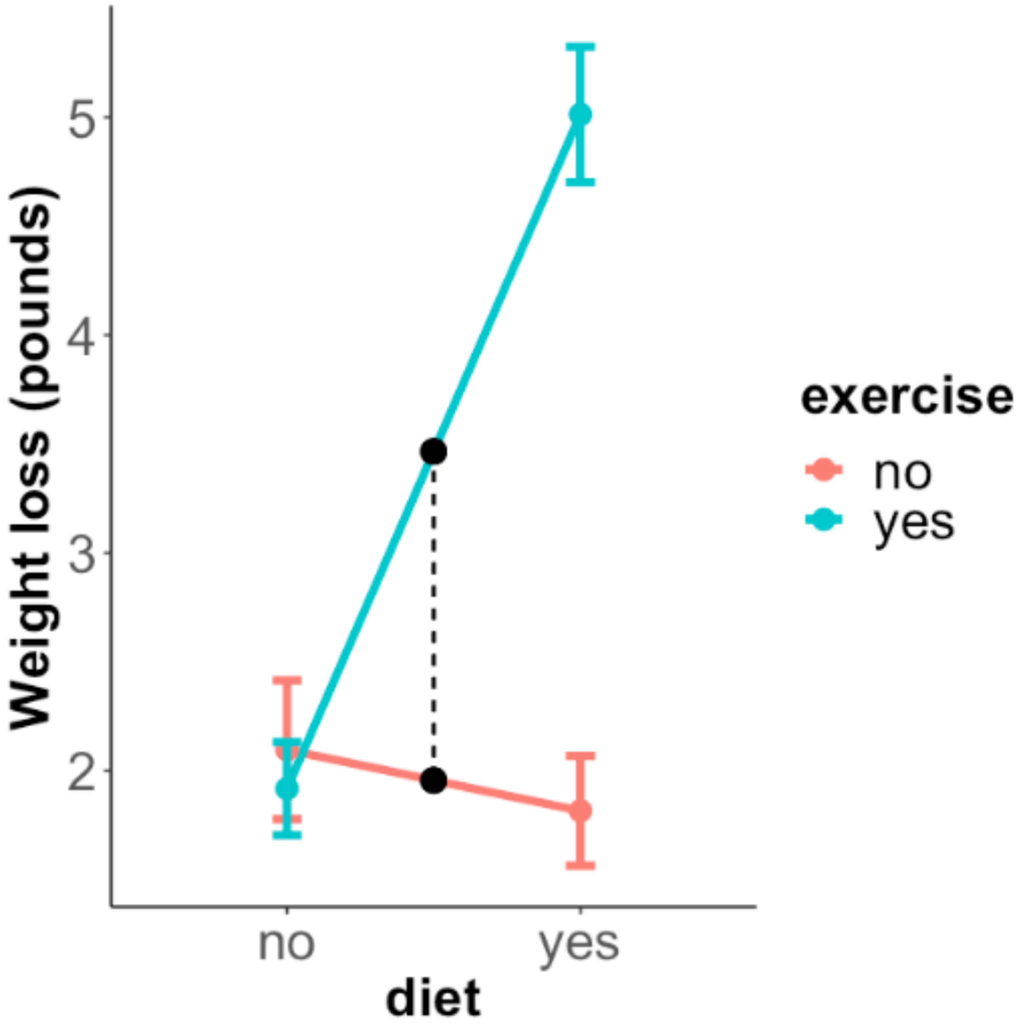
CASE 4.1: only interaction should be interpreted but not the main effect.



```
> anova(lm(WeightLoss~Diet*Exercise))
Analysis of Variance Table

Response: WeightLoss
              Df  Sum Sq Mean Sq F value    Pr(>F)
Diet           1  9.9152  9.9152  98.969 2.952e-08 ***
Exercise       1 11.4031 11.4031 113.820 1.108e-08 ***
Diet:Exercise  1 14.2526 14.2526 142.263 2.246e-09 ***
Residuals     16  1.6030  0.1002
```

CASE 4: One or both main factors are significant and the interaction as well. CASE 4.1: only interaction should be interpreted but not the main effect.



A main effect says that there is a difference in weight loss between the exercise means, regardless of diet.

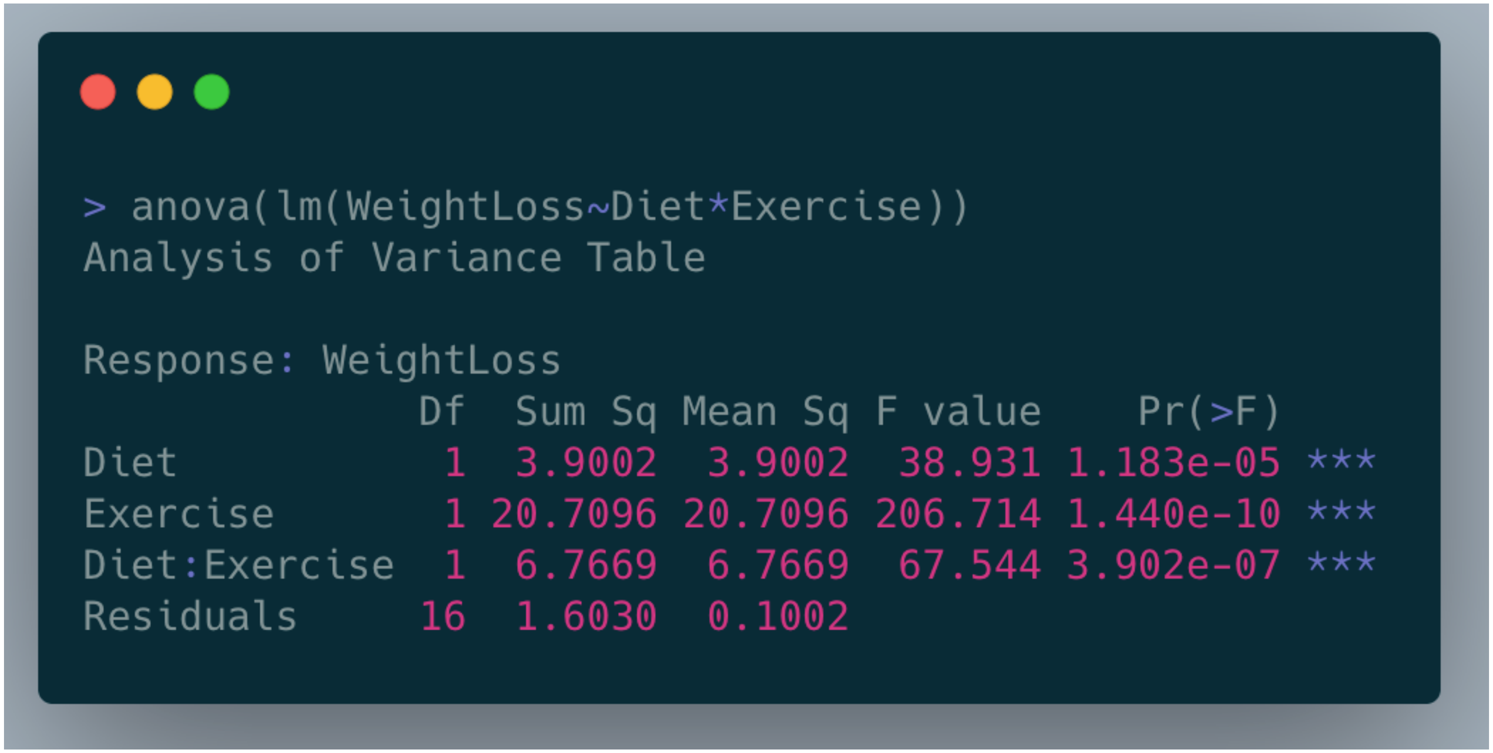
This may be technically true but only because of the big differences in weight loss due to diet.

It is not true that weight loss differs for the no diet case.

So, to say that there is weight loss regardless of diet (main effect) is not accurate!

CASE 4: One or both main factors are significant and the interaction as well.

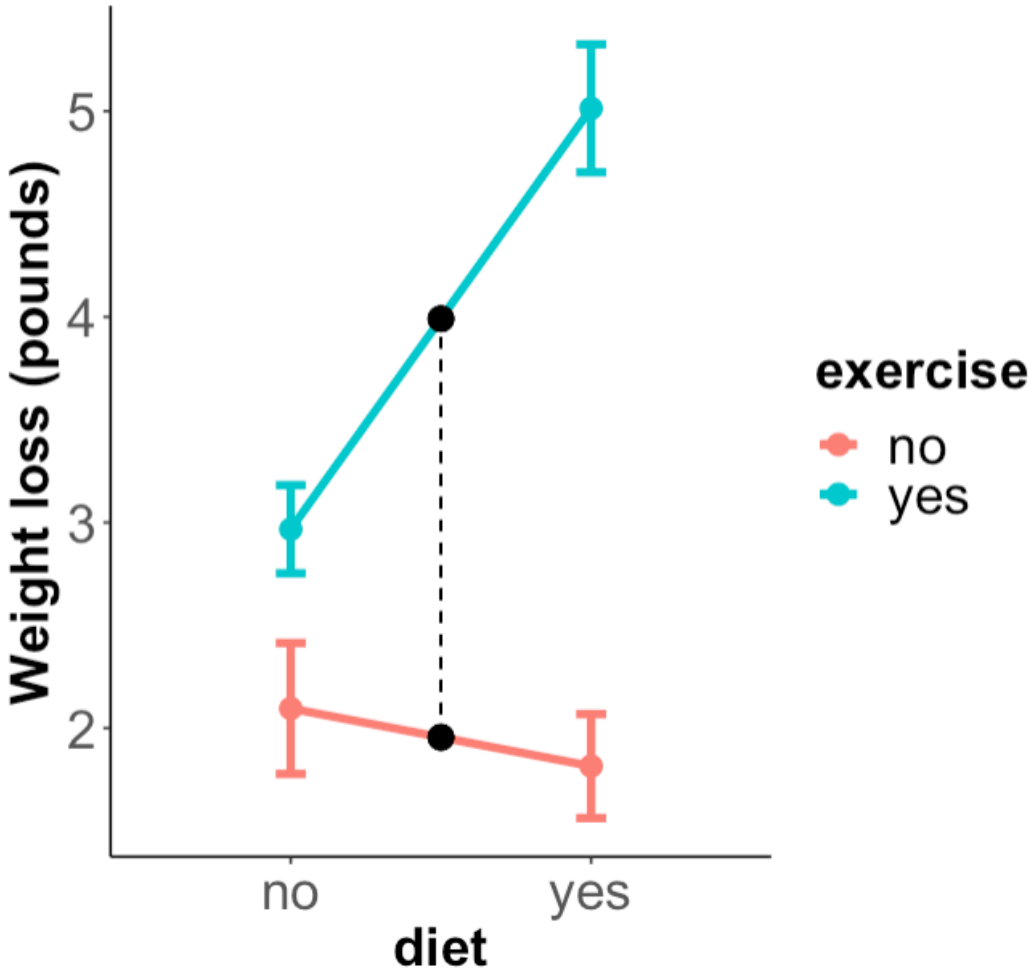
CASE 4.2: the interaction & main effect can be interpreted.



```
> anova(lm(WeightLoss~Diet*Exercise))
Analysis of Variance Table

Response: WeightLoss
              Df  Sum Sq Mean Sq F value    Pr(>F)    
Diet           1  3.9002  3.9002  38.931 1.183e-05 ***
Exercise       1 20.7096 20.7096 206.714 1.440e-10 ***
Diet:Exercise  1  6.7669  6.7669  67.544 3.902e-07 ***
Residuals     16  1.6030  0.1002                      
```

CASE 4: One or both main factors are significant and the interaction as well. CASE 4.2: the interaction & main effect can be interpreted.



A main effect says that there is a difference in weight loss between the exercise means, regardless of diet.

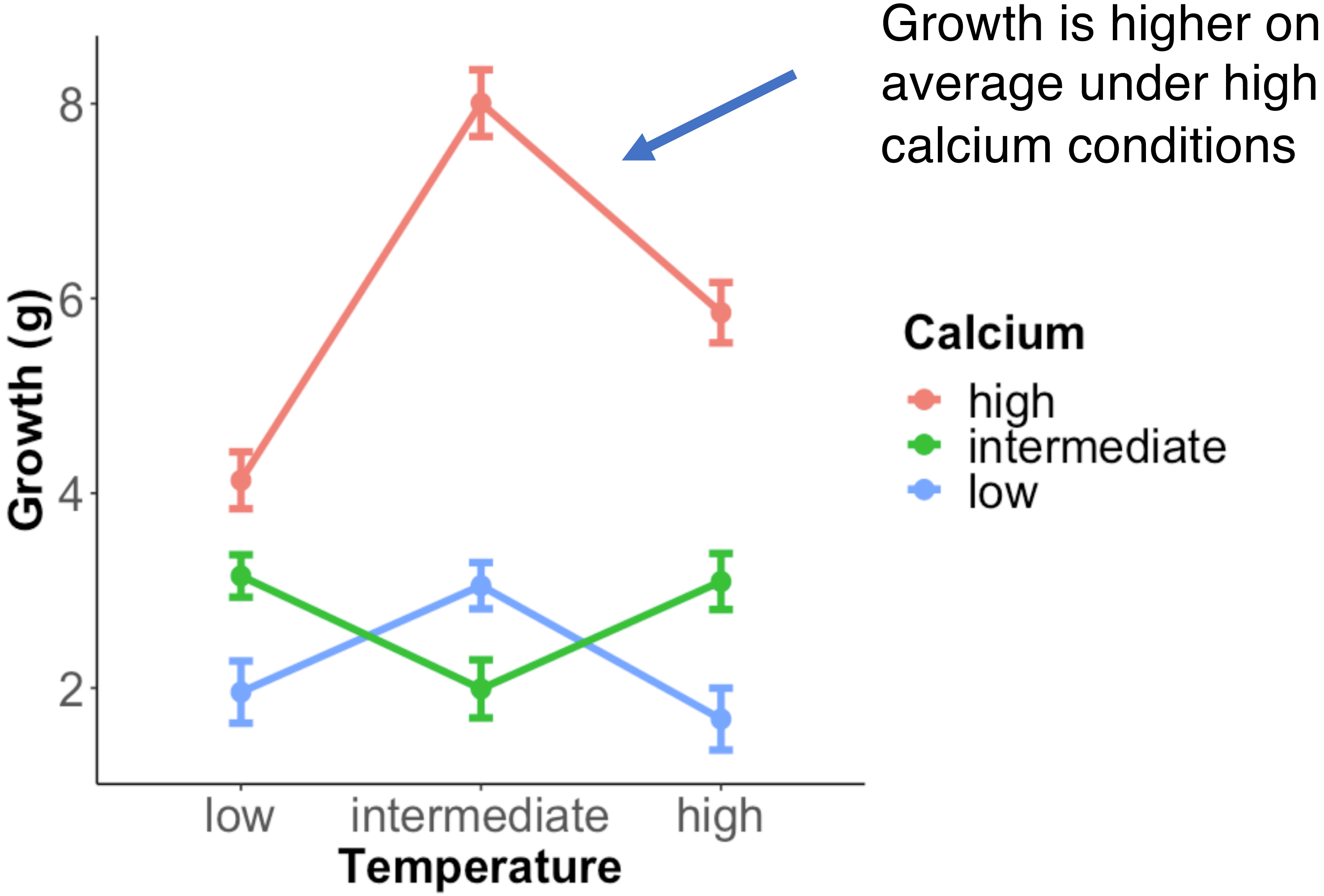
This is the case here because the weight loss when individuals exercised is consistently greater than no exercise regardless of the diet.

And individuals that exercised and dieted loss even more weight than individuals than only dieted.

Why can we interpret both the interaction and the main effect?

One could interpret both the interaction and the main effect because the interaction does not obscure a consistent overall pattern.

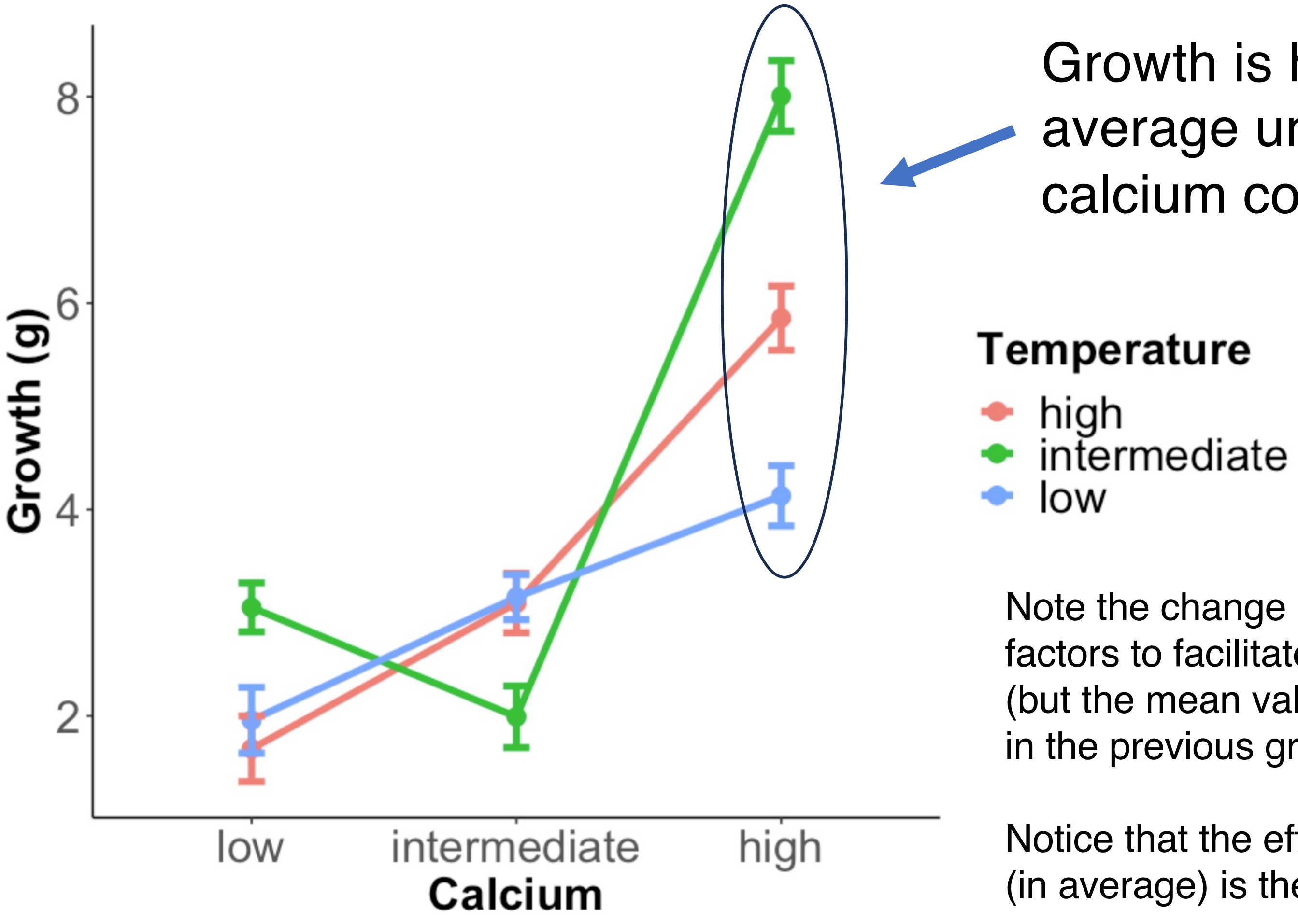
Although growth clearly depends on the combination of temperature and calcium (non-parallel lines → interaction), one calcium level (high) shows consistently higher growth than the others across all temperatures. That consistency means the average effect of calcium remains meaningful, even in the presence of an interaction.



Growth is higher on average under high calcium conditions

Why should we focus on the interaction only?

In this plot, the effect of calcium on growth strongly depends on temperature, and—critically—the direction of the effect changes across temperatures. At low and high temperatures, growth increases with calcium, whereas at intermediate temperature the pattern is reversed at low to intermediate calcium before sharply increasing. As a result, the ranking of calcium levels is not consistent across temperatures, and neither is the ranking of temperature levels across calcium concentrations.



Growth is higher on average under high calcium conditions

Note the change in the position of factors to facilitate the interpretation (but the mean values are the same as in the previous graph).

Notice that the effect of high Calcium (in average) is the same.

Multi-factorial ANOVA

Assumptions (the same as for the one-way ANOVA):

1) Each of the samples (observations within groups) is a random sample from its population (LATER IN THE COURSE).

2) The variable (e.g., weight loss) is normally distributed in each combination of treatment (e.g., no diet and exercise) population.

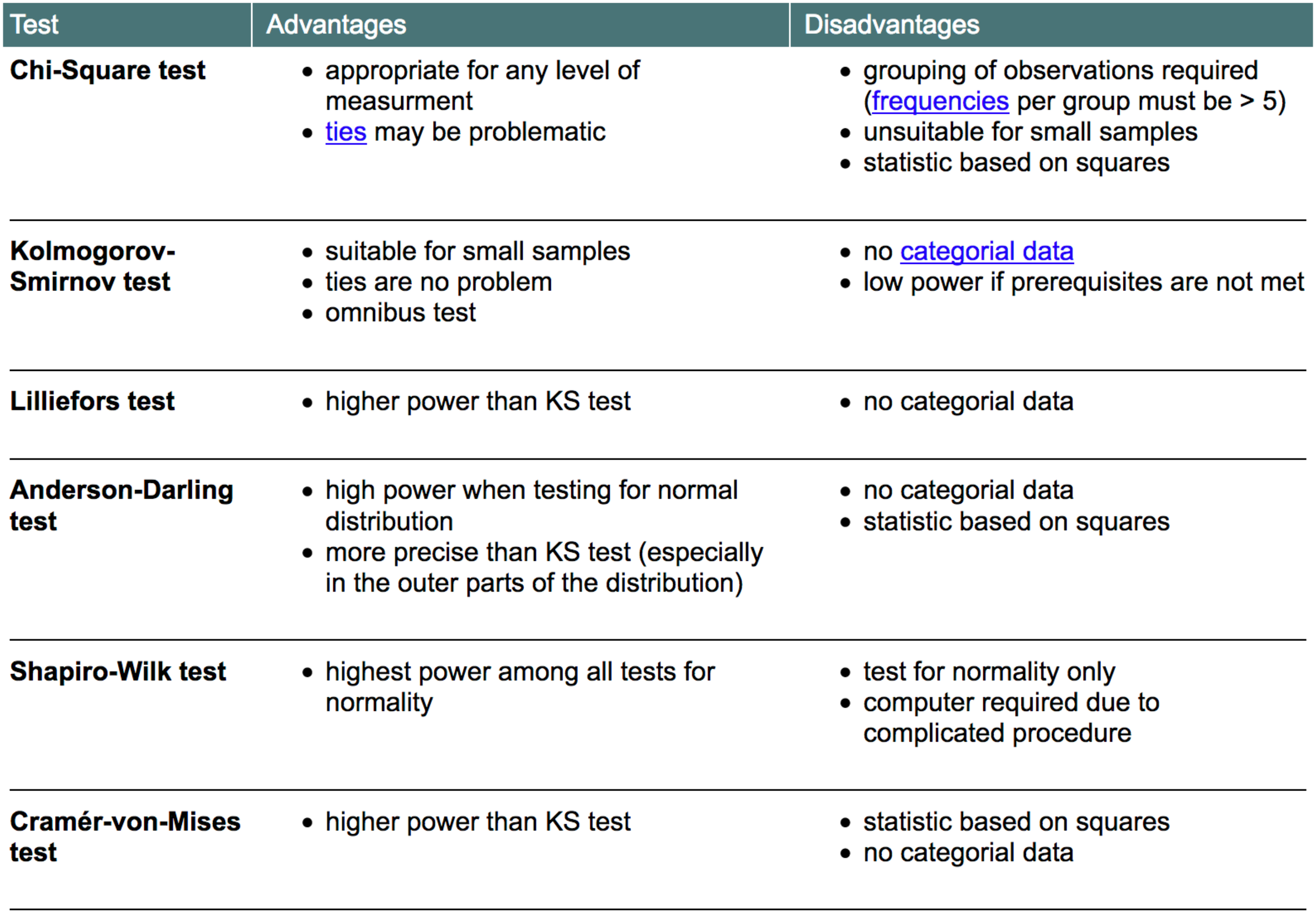
3) The variances are equal among all populations from which the treatments were sampled (otherwise the F values change in ways that may not measure difference among means).

Assessing the normality assumption

- ANOVAs are not very sensitive to lack of normality (i.e., they are robust against normality).

- Simulation studies, using a variety of non-normal distributions, have shown that the false positive rates (Type I error rates) in ANOVA are not strongly affected by the violation of the normality assumption (Harwell et al. 1992, Lix et al. 1996).

Assessing the normality assumption – some traditional tests



Test	Advantages	Disadvantages
Chi-Square test	• appropriate for any level of measurment • ties may be problematic	• grouping of observations required (frequencies per group must be > 5) • unsuitable for small samples • statistic based on squares
Kolmogorov-Smirnov test	• suitable for small samples • ties are no problem • omnibus test	• no categorial data • low power if prerequisites are not met
Lilliefors test	• higher power than KS test	• no categorial data
Anderson-Darling test	• high power when testing for normal distribution • more precise than KS test (especially in the outer parts of the distribution)	• no categorial data • statistic based on squares
Shapiro-Wilk test	• highest power among all tests for normality	• test for normality only • computer required due to complicated procedure
Cramér-von-Mises test	• higher power than KS test	• statistic based on squares • no categorial data

Source: http://www.statistics4u.info/fundstat_eng/cc_normality_test.html

Assessing the normality assumption:
The Quantile-Quantile normal plot (Q-Q normal plot)

The normal Q-Q plot is a graphical technique for determining if multiple data sets come from populations with a common distribution (here, if they all come from normally distributed populations regardless of their means and variances).

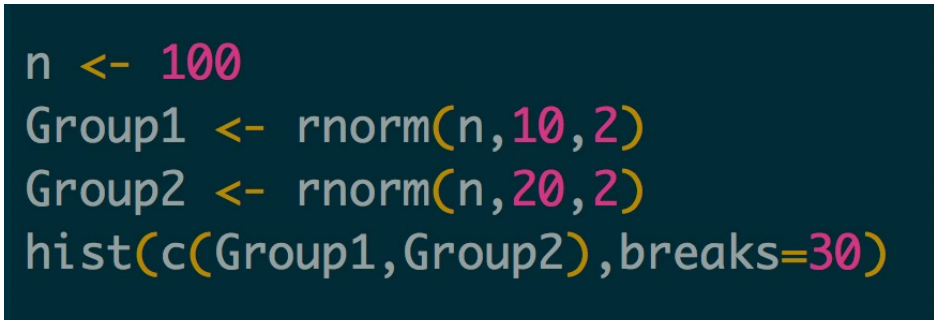
Tutorial 3: Factorial ANOVA

Factorial Analysis of Variance

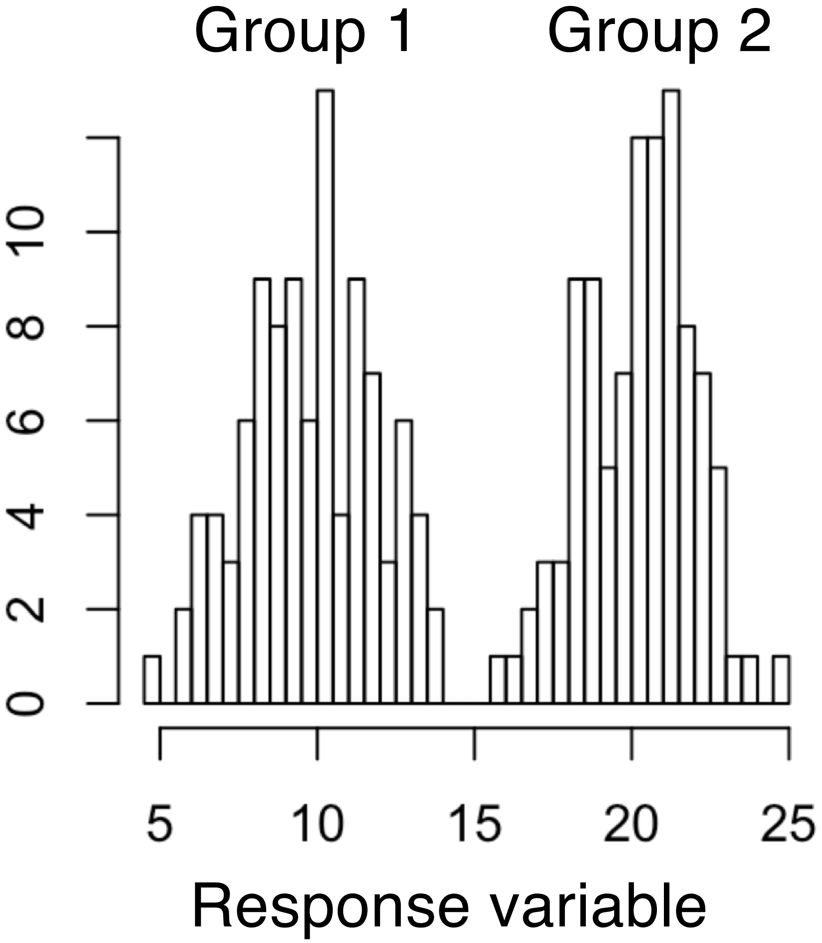
Assessing the normality assumption in linear models: The Quantile-Quantile normal (Q-Q normal plot)

In ANOVA, normality is not required for the response variable overall (e.g., weight loss), but for the distribution of the response within each group.

Usual interpretation of the normality assumption in ANOVAs - "Data have to be normal"

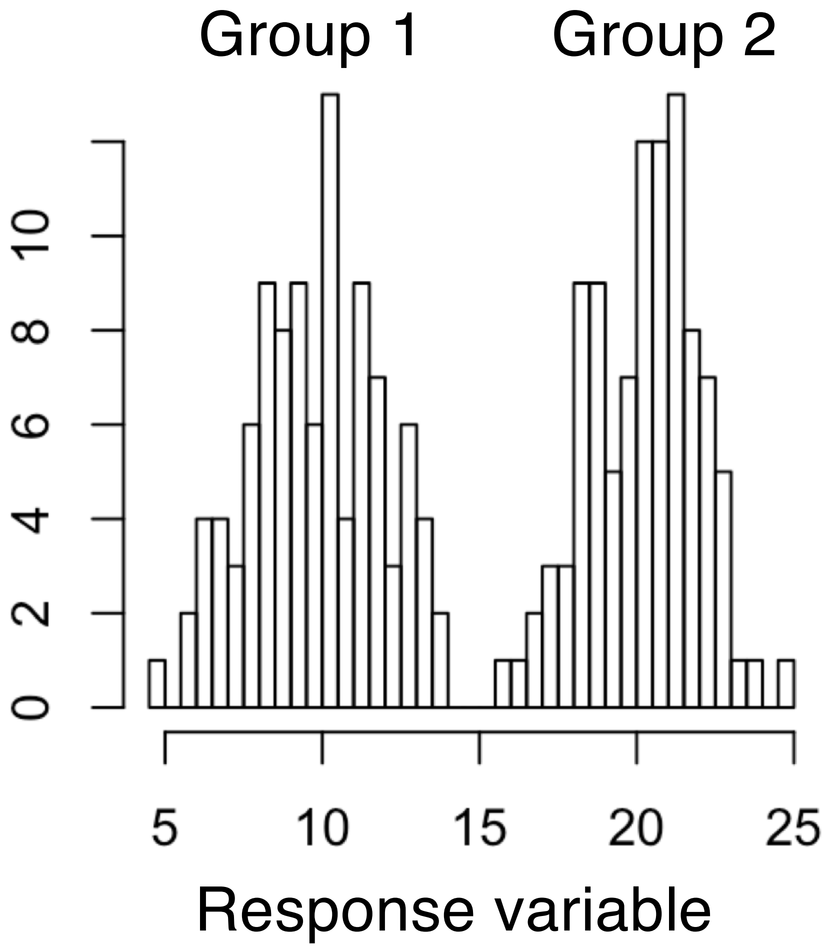


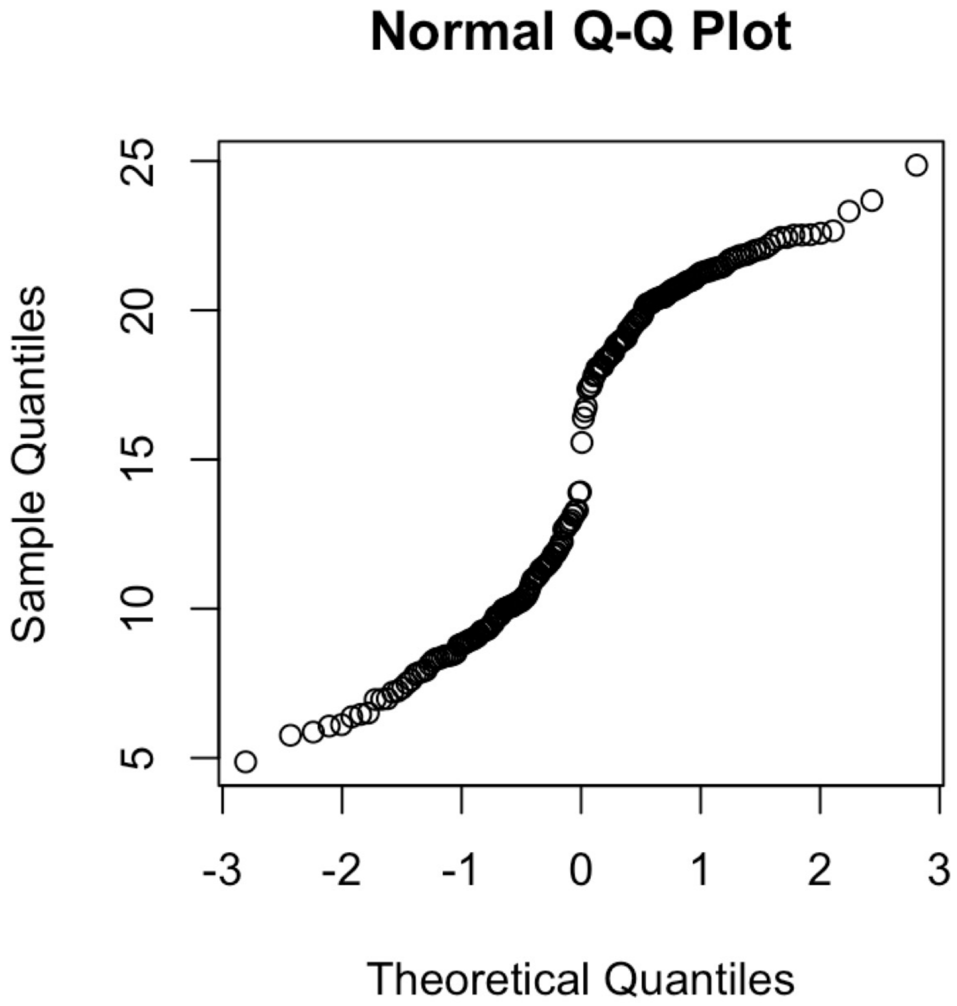
```
n <- 100
Group1 <- rnorm(n,10,2)
Group2 <- rnorm(n,20,2)
hist(c(Group1,Group2),breaks=30)
```



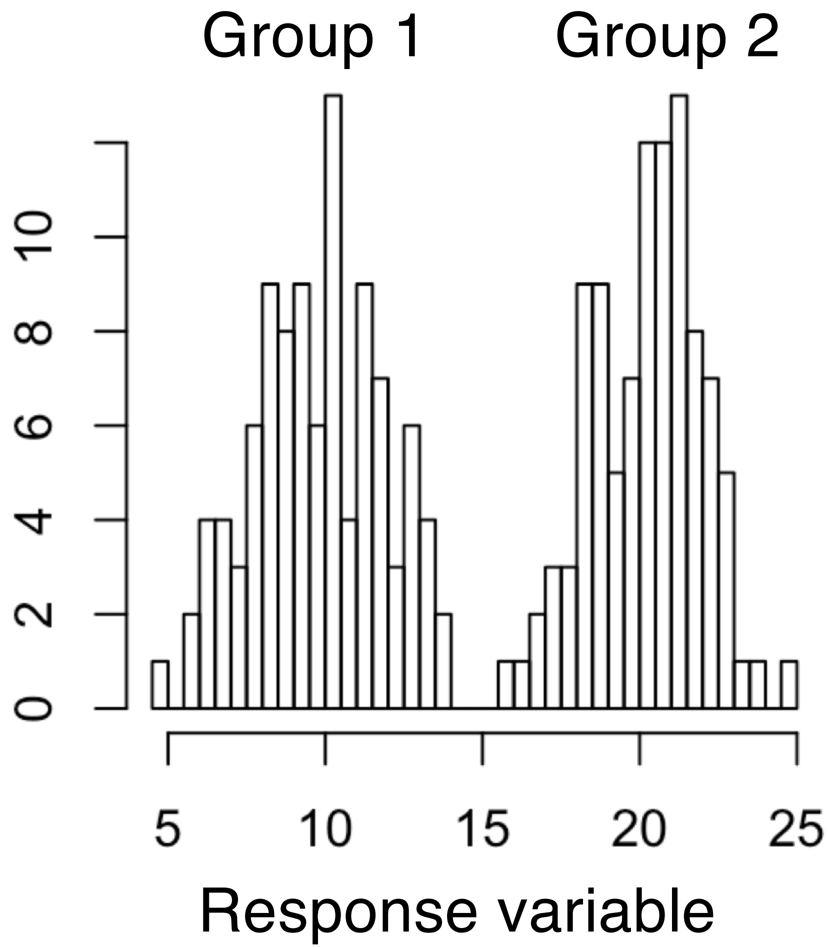
Response variable not normal across groups, but normal within groups (the correct assumption).

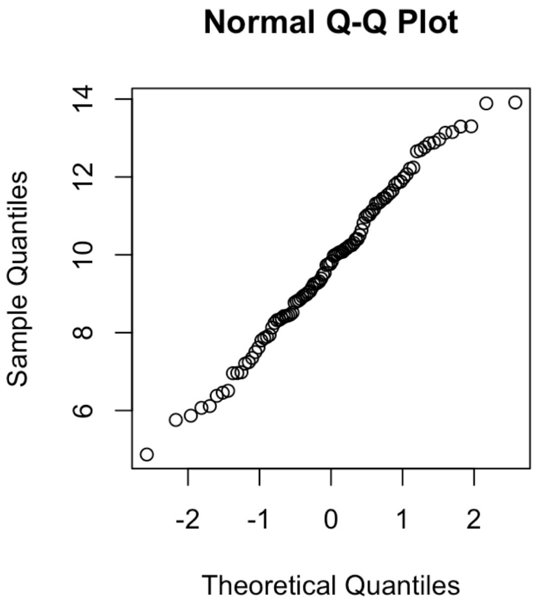
Assessing the normality assumption in linear models: The Quantile-Quantile normal (Q-Q normal plot)



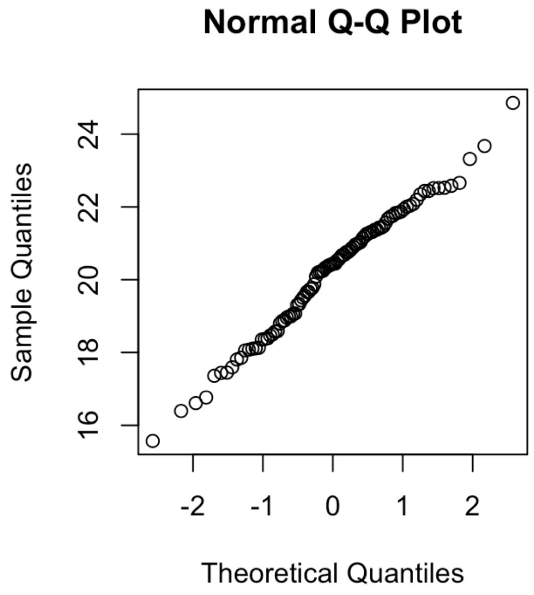


Assessing the normality assumption in linear models: The Quantile-Quantile normal (Q-Q normal plot)





Group 1

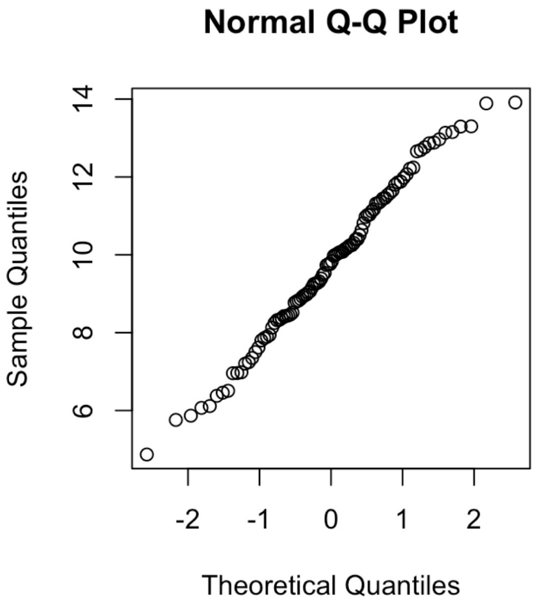


Group 2

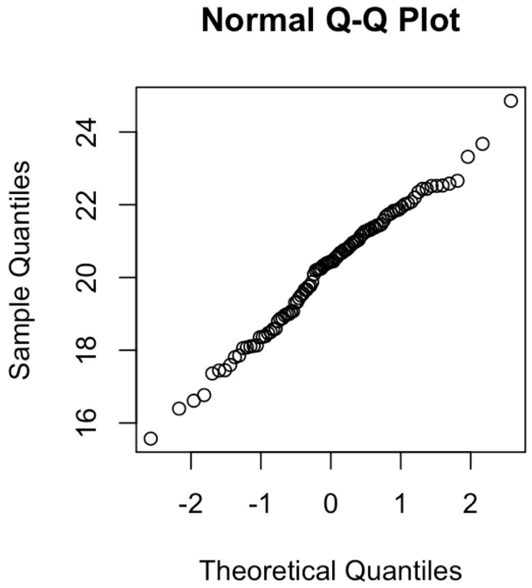
Assessing the normality assumption in linear models: The Quantile-Quantile normal (Q-Q normal plot)

When there are many factors and levels (e.g., multi-factorial ANOVA), checking normality with Q–Q plots for every group quickly becomes impractical.

Example: 2 factors with 3 and 4 levels already produce 12 groups.



Group 1



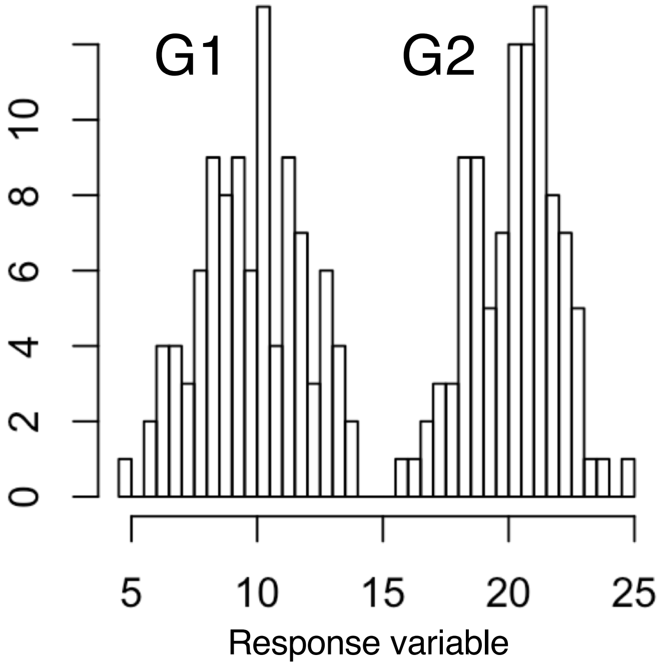
Group 2

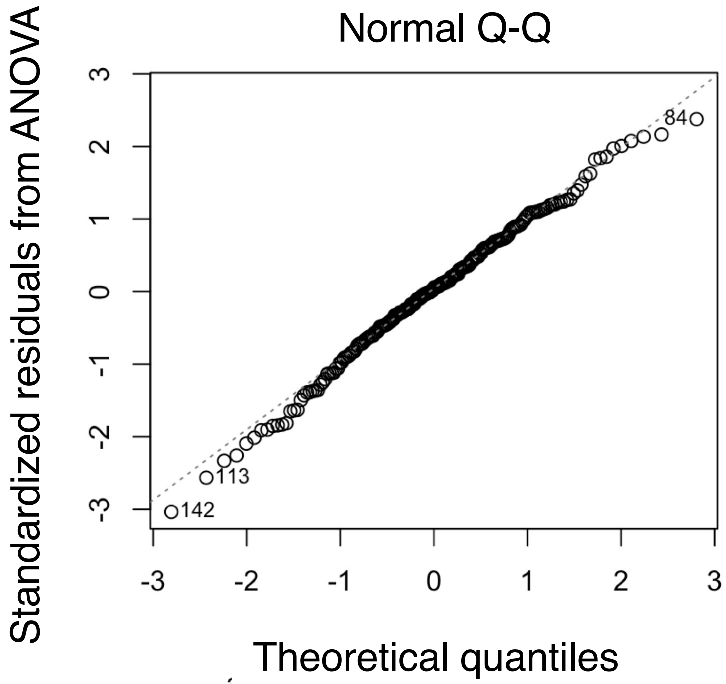
Assessing the normality assumption in linear models:
The Quantile-Quantile normal plot of residuals (Q-Q normal residual plot)

ANOVA is a linear multiple regression model in which the response variable is continuous, and predictors are categorical.

$$Y = Factor(G1, G2) + residuals$$

So, instead of plotting all groups, we assess the residuals across all groups, i.e., variation not accounted by group mean differences.





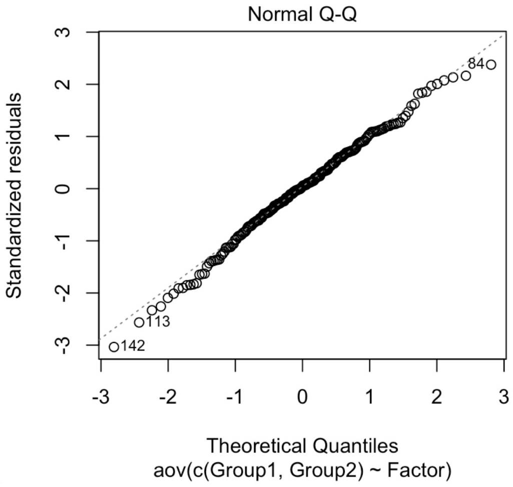
Assessing the normality assumption in linear models: The Quantile-Quantile normal plot of residuals (Q-Q normal residual plot)

You will practice the application of Q-Q normal residual plots for two-factorial ANOVAs in tutorial 3.

$$Weight_{Loss} = Diet + Exercise + Diet \times Exercise + residuals$$

Tutorial 3: Factorial ANOVA

Factorial Analysis of Variance



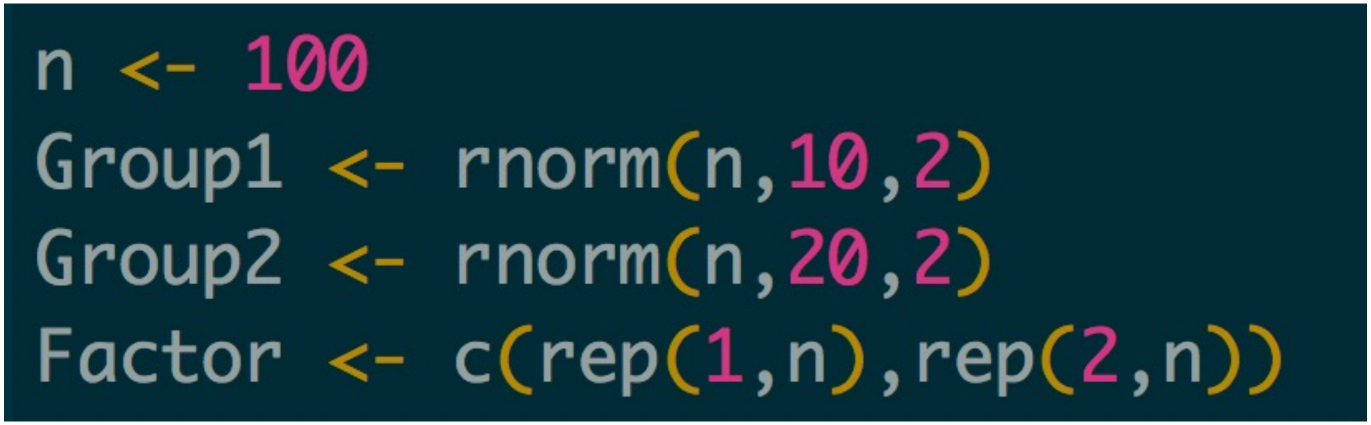
Assessing the equality of variance (homoscedasticity) assumption

Two commonly used tests for the null hypothesis that multiple samples come from populations with equal variances are Levene's test and Bartlett's test (the latter is more sensitive to departures from normality).

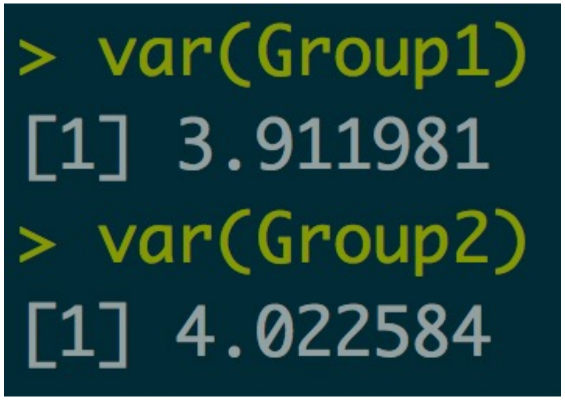
H_0: All groups come from populations with equal variances.
H_A: At least one group comes from a population with a different variance.

We say that groups come from populations with equal variances, not from the same population, because populations can share the same variance while having different means (and vice versa). In other words, populations can be different even when one of their properties is the same.

Assessing the equality of variance (homoscedasticity) assumption

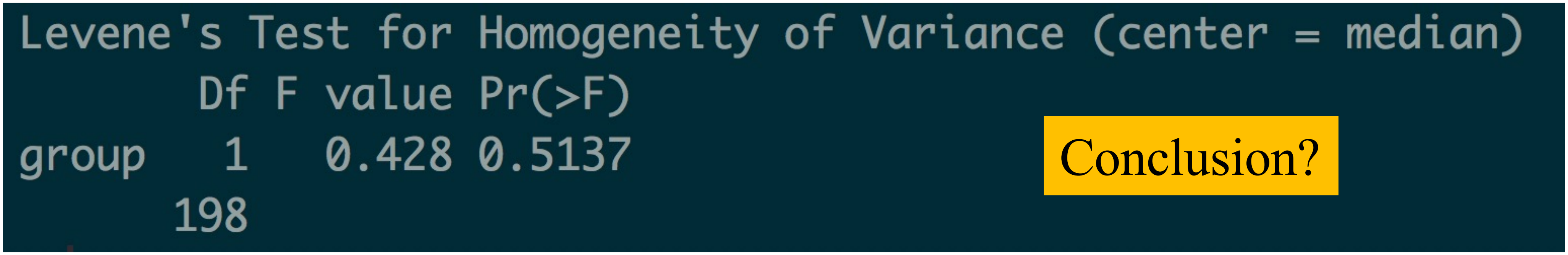


```
n <- 100
Group1 <- rnorm(n,10,2)
Group2 <- rnorm(n,20,2)
Factor <- c(rep(1,n),rep(2,n))
```



```
> var(Group1)
[1] 3.911981
> var(Group2)
[1] 4.022584
```

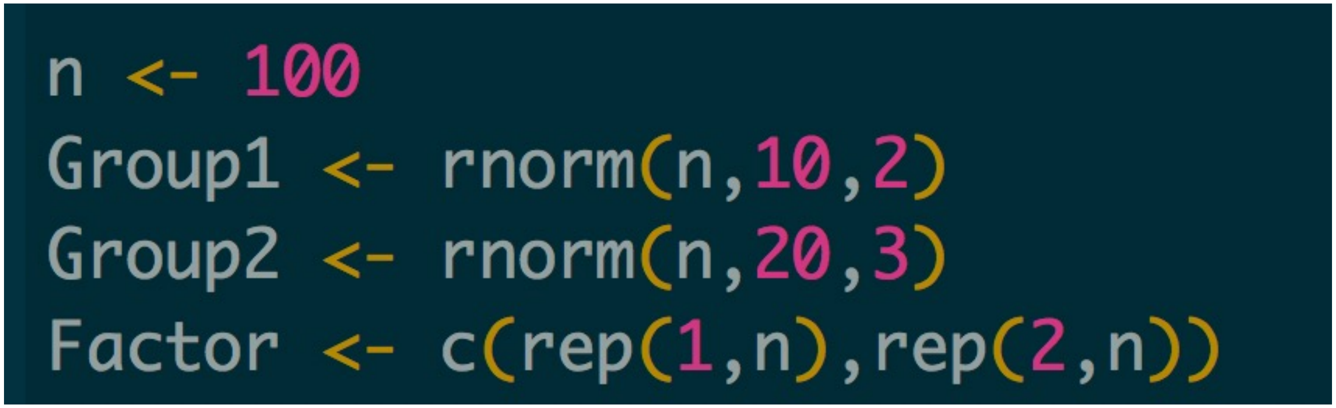
The two samples come from populations with the same variances (they only vary in mean values).



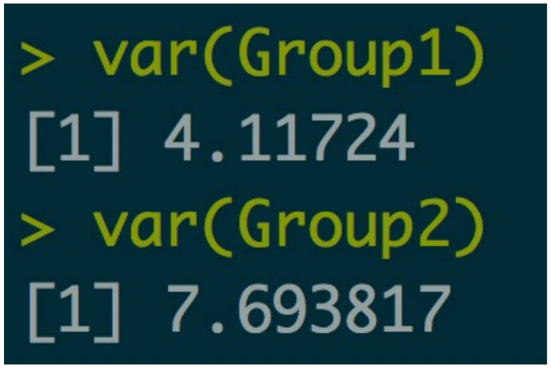
```
Levene's Test for Homogeneity of Variance (center = median)
       Df F value Pr(>F)
group   1   0.428 0.5137
      198
```

Conclusion?

Assessing the equality of variance (homoscedasticity) assumption

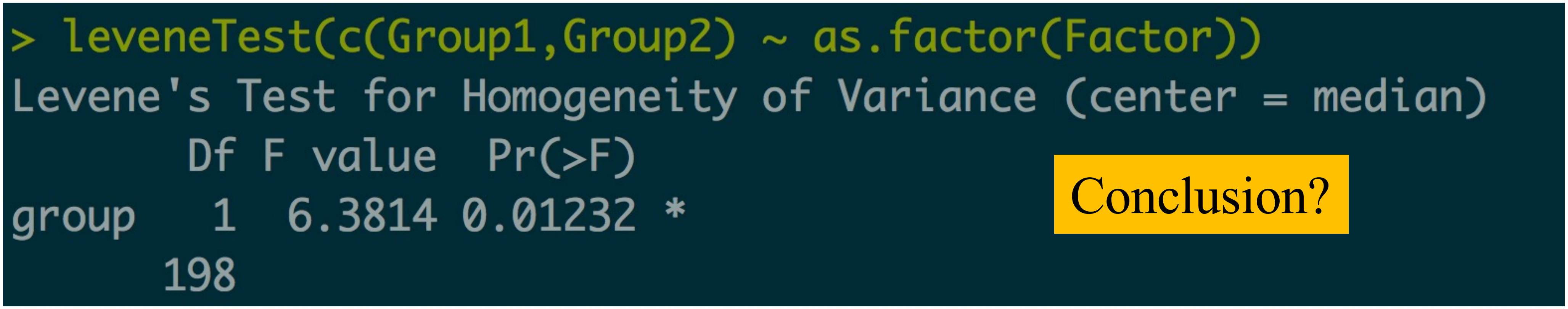


```
n <- 100
Group1 <- rnorm(n,10,2)
Group2 <- rnorm(n,20,3)
Factor <- c(rep(1,n),rep(2,n))
```



```
> var(Group1)
[1] 4.11724
> var(Group2)
[1] 7.693817
```

The two samples come from populations with different variances (and they also vary in their means).



```
> leveneTest(c(Group1,Group2) ~ as.factor(Factor))
Levene's Test for Homogeneity of Variance (center = median)
       Df F value  Pr(>F)  
group   1  6.3814 0.01232 *
      198                  
```

Conclusion?

Let’s contrast the Levene’s and ANOVAs hypotheses

Levene’s:

H_0: All groups come from populations with equal variances.

H_A: At least one group comes from a population with a different variance.

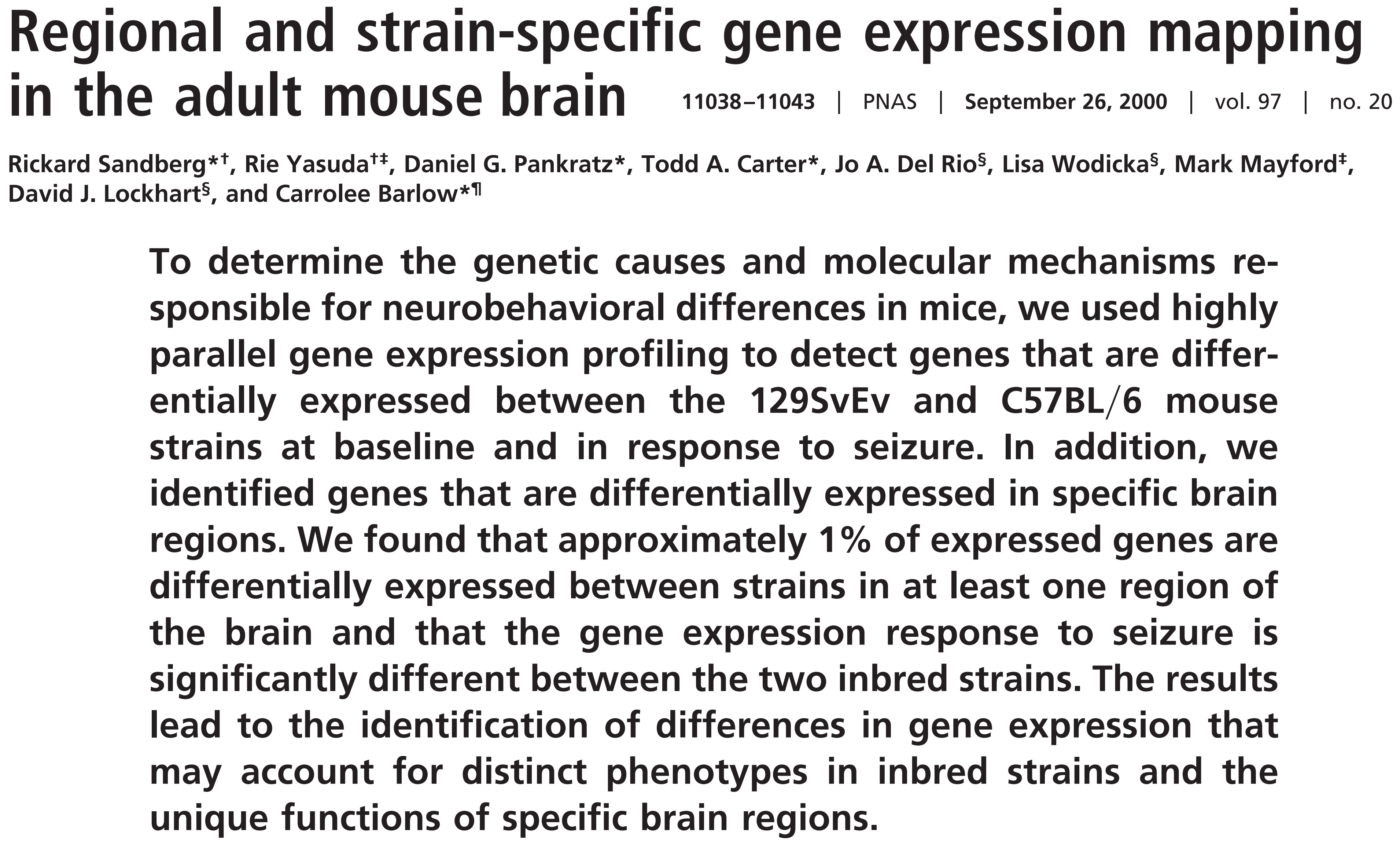
ANOVA:

H_0: The samples come from the same population.
H_A: At least two samples come from different populations.

If populations are normally distributed and share the same variance and the same mean, then under the null hypothesis they can be considered as coming from the same population.

This is why we first test for equality of variances (e.g., using Levene’s test) before conducting an ANOVA.

A more complex (and real) biological data



Regional and strain-specific gene expression mapping in the adult mouse brain

11038–11043 | PNAS | September 26, 2000 | vol. 97 | no. 20

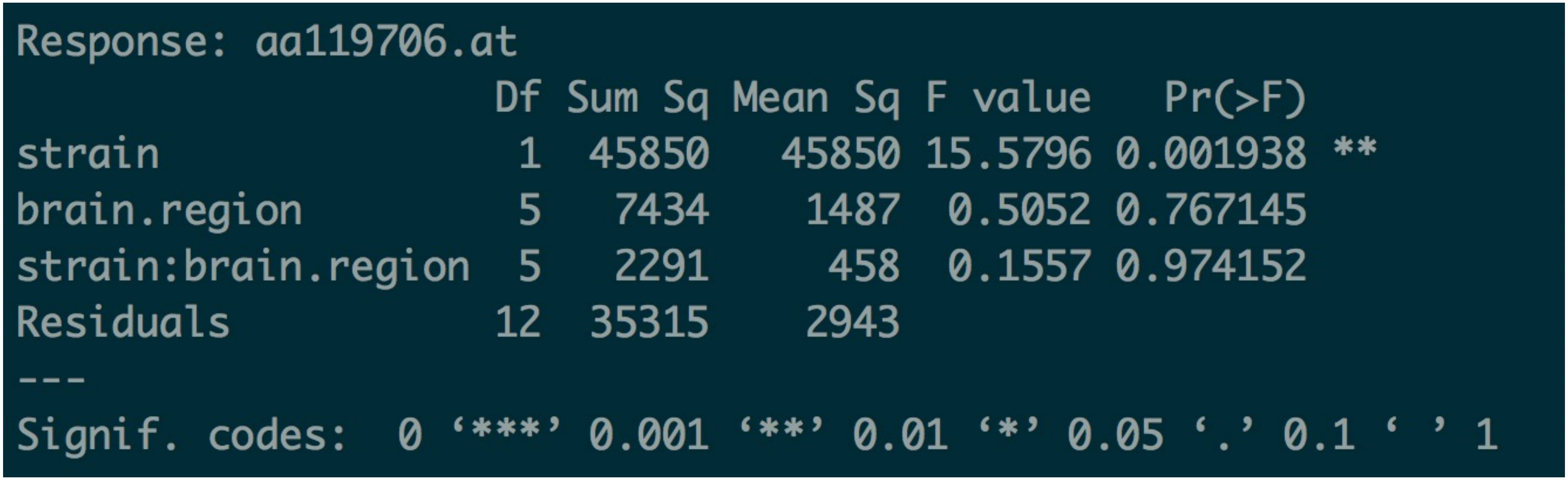
Rickard Sandberg[*†], Rie Yasuda[†‡], Daniel G. Pankratz[*], Todd A. Carter[*], Jo A. Del Rio[§], Lisa Wodicka[§], Mark Mayford[‡], David J. Lockhart[§], and Carrolee Barlow[*¶]

To determine the genetic causes and molecular mechanisms responsible for neurobehavioral differences in mice, we used highly parallel gene expression profiling to detect genes that are differentially expressed between the 129SvEv and C57BL/6 mouse strains at baseline and in response to seizure. In addition, we identified genes that are differentially expressed in specific brain regions. We found that approximately 1% of expressed genes are differentially expressed between strains in at least one region of the brain and that the gene expression response to seizure is significantly different between the two inbred strains. The results lead to the identification of differences in gene expression that may account for distinct phenotypes in inbred strains and the unique functions of specific brain regions.

Gene expression is standardized in relation to seizure versus base line

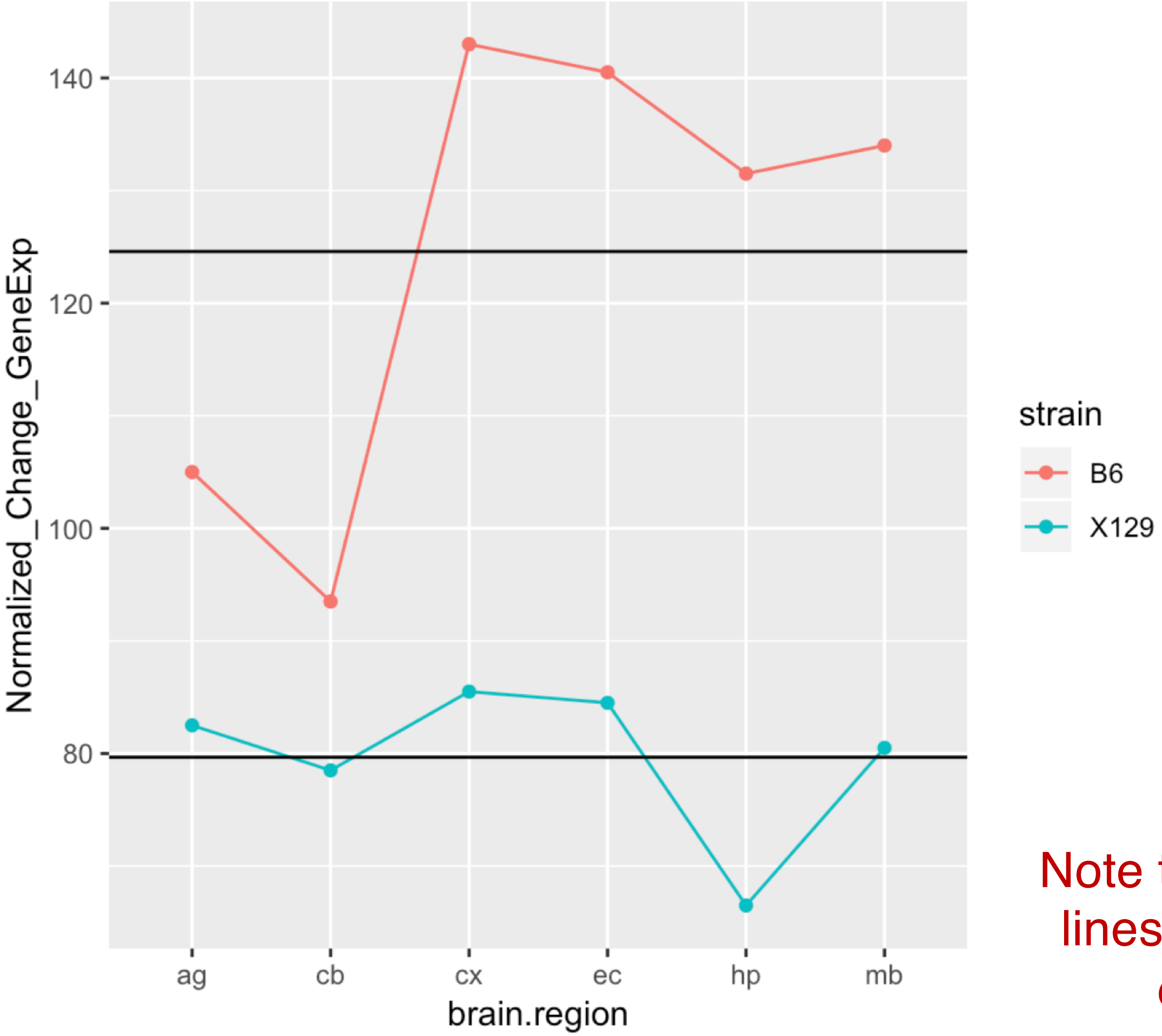
Case 1 - What are the significant effects?
a gene for which only strain is significant (i.e., they differ in gene expression levels)

gene aa119706.at - Only strain is significant (i.e., strains differ from one another in their mean gene expression levels, but these differences are independent of the brain region)



```
Response: aa119706.at
                  Df Sum Sq Mean Sq F value   Pr(>F)   
strain             1  45850   45850 15.5796 0.001938 **
brain.region       5   7434    1487  0.5052 0.767145   
strain:brain.region  5   2291     458  0.1557 0.974152   
Residuals         12  35315    2943                    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

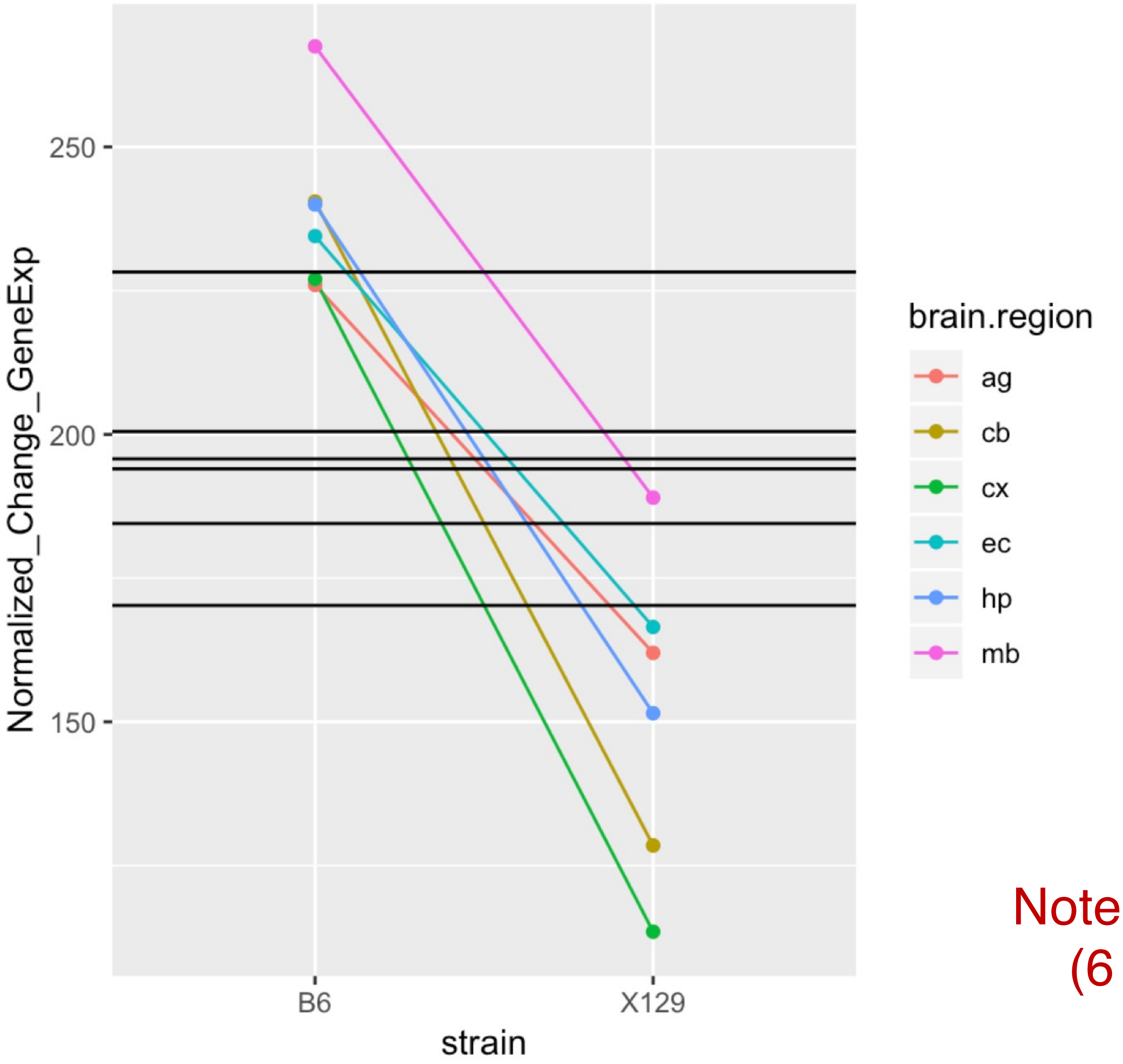
gene aa217379.s.at - Only strain is significant



Black lines represent means of each strain independent of brain region and their difference is significant.

Note that the B6 and X129 lines (orange and green) do look parallel

gene aa217379.s.at - Only strain is significant

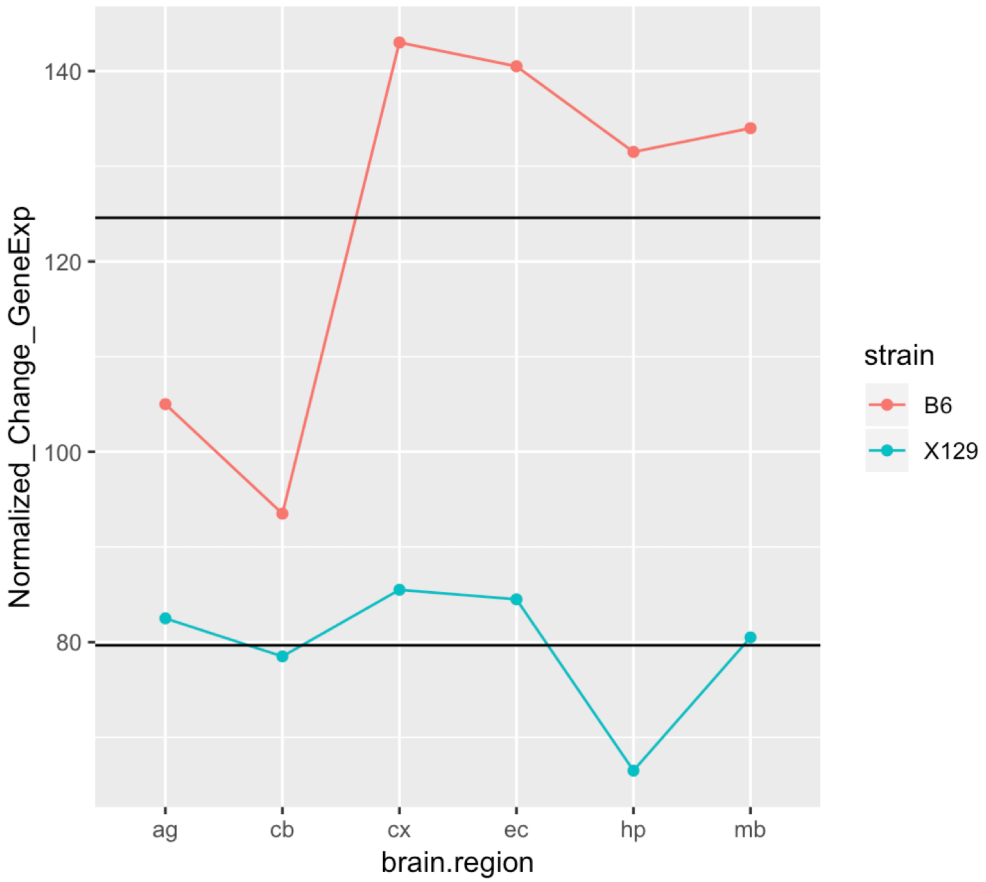


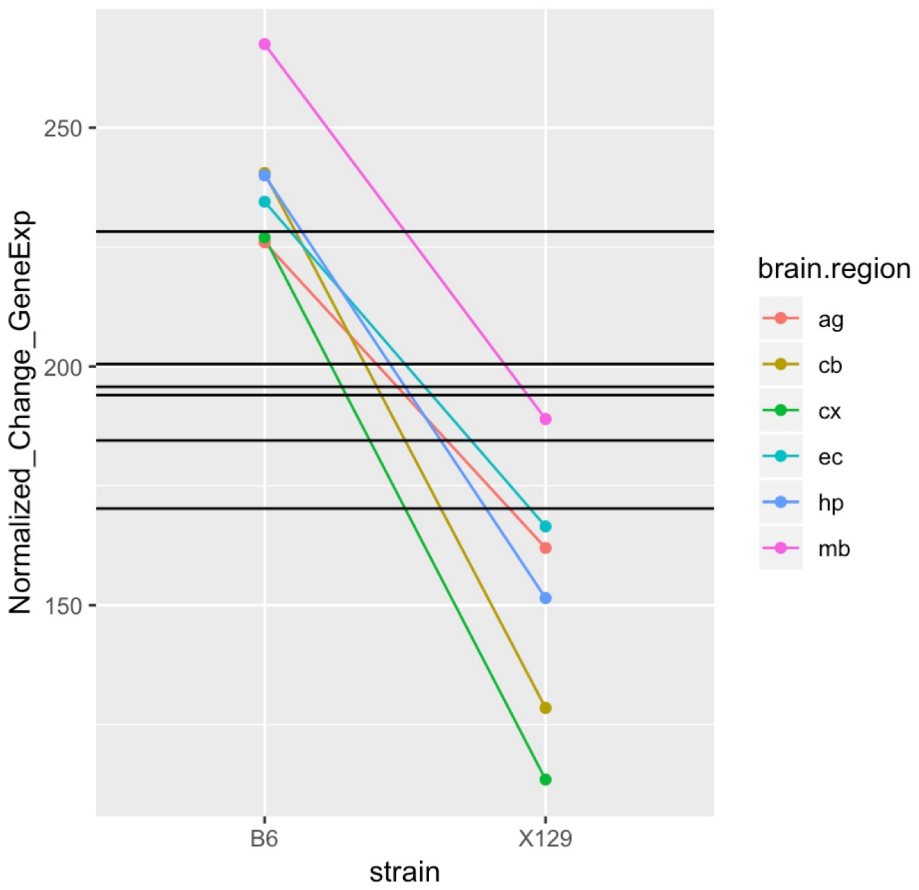
Black lines represent means of each brain region independent of strain and their differences are not significant.

Note that the brain region lines (6 colors) do look parallel

Which factor to plot where?

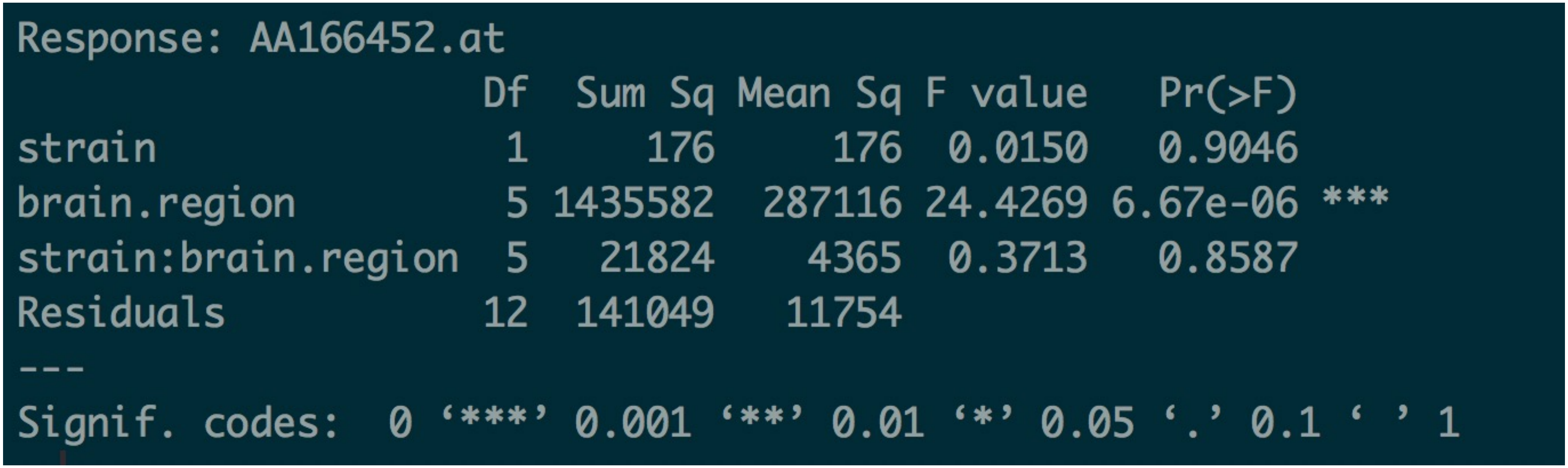
It depends on how differences facilitate interpretation.





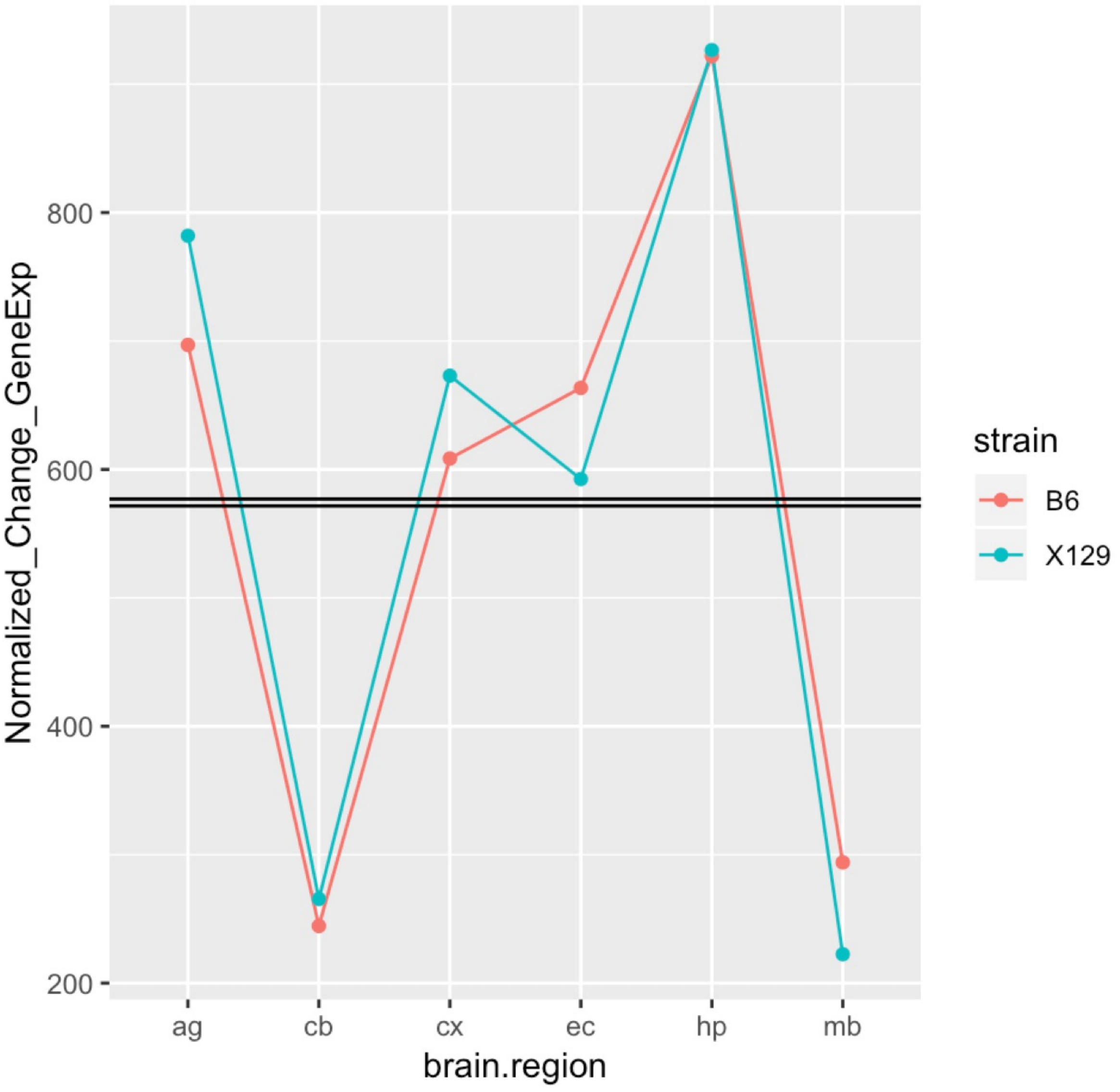
Case 2 - What are the significant effects?
a gene for which only the brain region is significant

gene AA166452.at - Only brain region is significant (i.e., regions differ from one another in their mean gene expression levels. but these differences are independent of the strain)



```
Response: AA166452.at
                   Df  Sum Sq Mean Sq F value    Pr(>F)    
strain              1     176     176  0.0150    0.9046    
brain.region        5 1435582  287116 24.4269 6.67e-06 ***
strain:brain.region 5   21824    4365  0.3713    0.8587    
Residuals          12  141049   11754                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

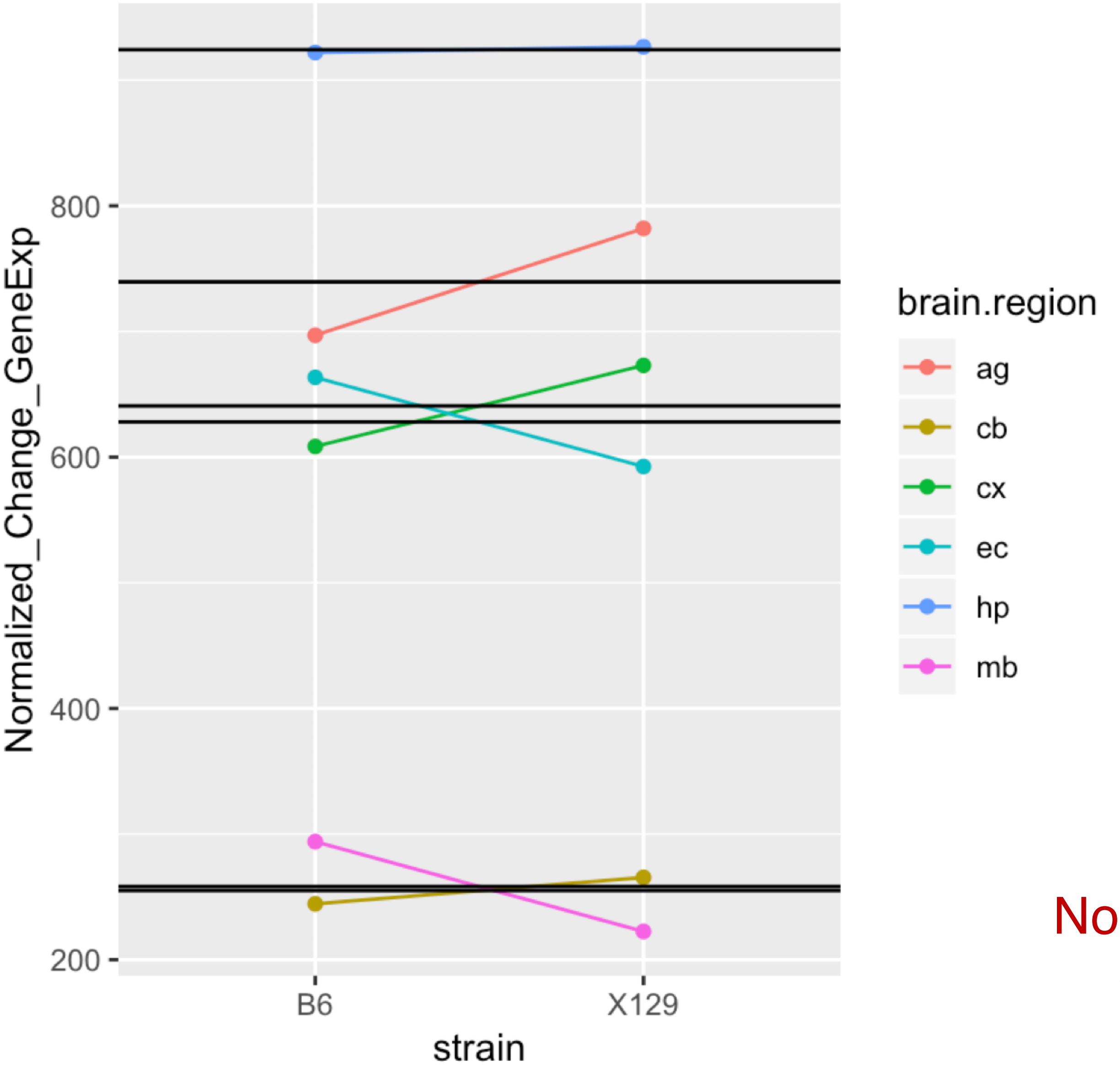
AA166452.at - Only brain region is significant



Black lines represent means of each strain independent of brain region and their difference are not significant.

Note that the B6 and X129 lines (orange and green) do look parallel

AA166452.at - Only brain region is significant

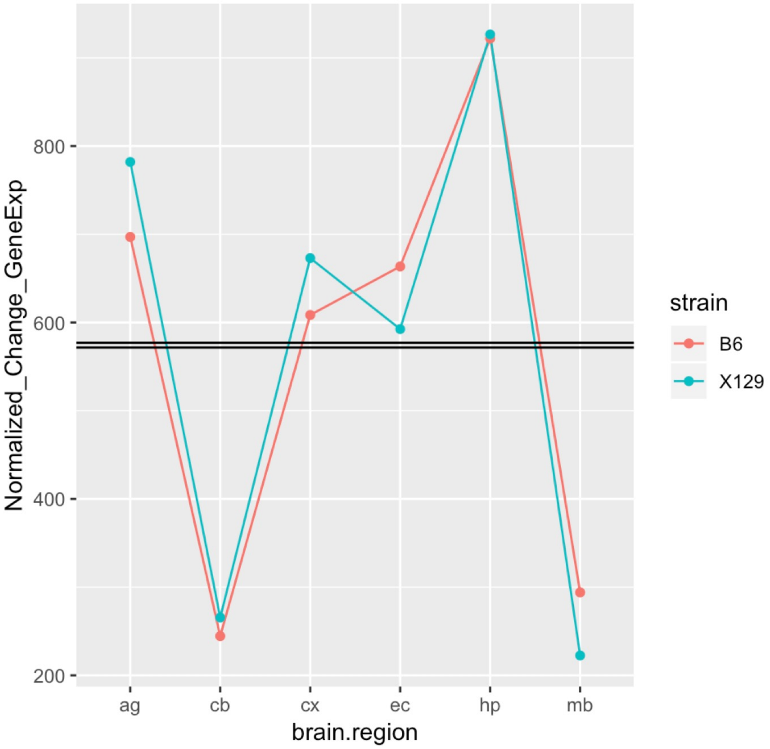


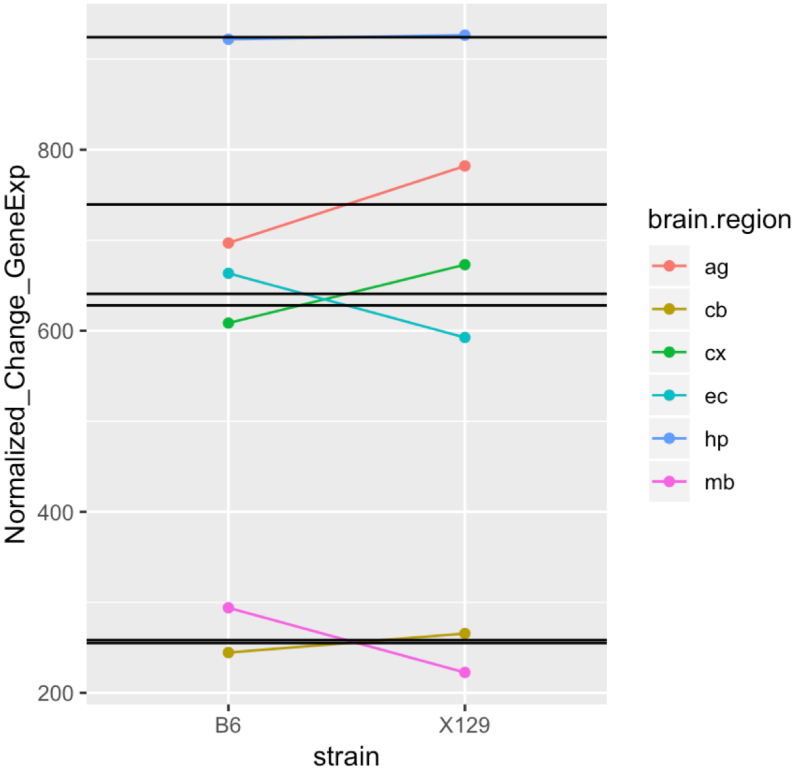
Black lines represent means of each brain region independent of strait and their differences are significant – BUT which brain regions differ from one another? There are 15 possible pairwise comparisons (tests).

Note that the brain region lines (6 colors) do look parallel

Which factor to plot where?

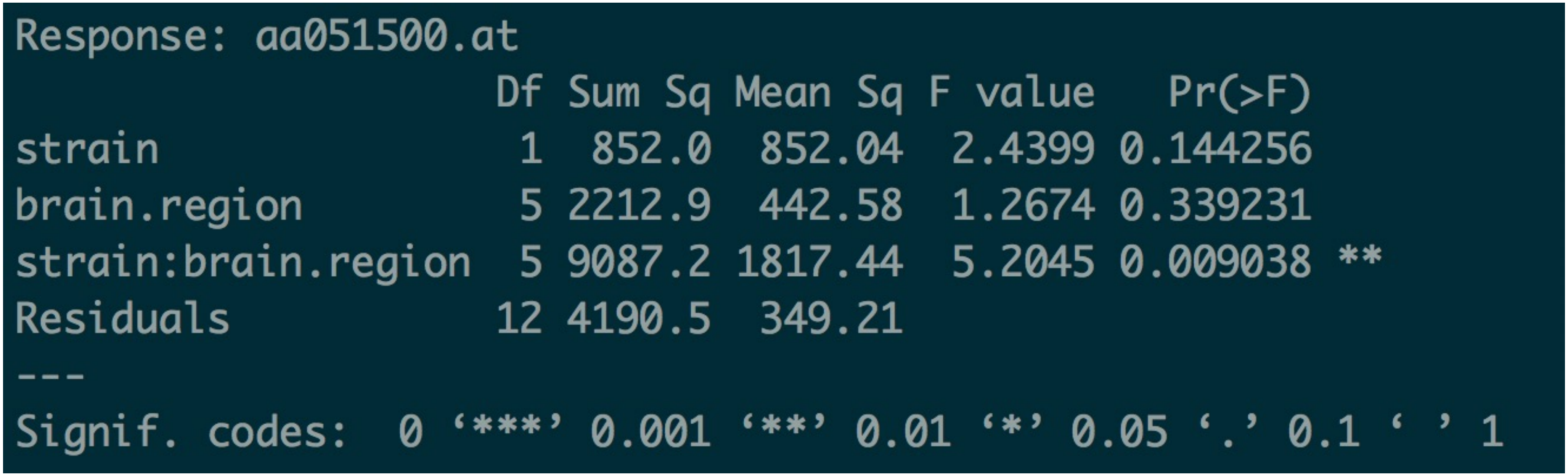
It depends on how differences facilitate interpretation





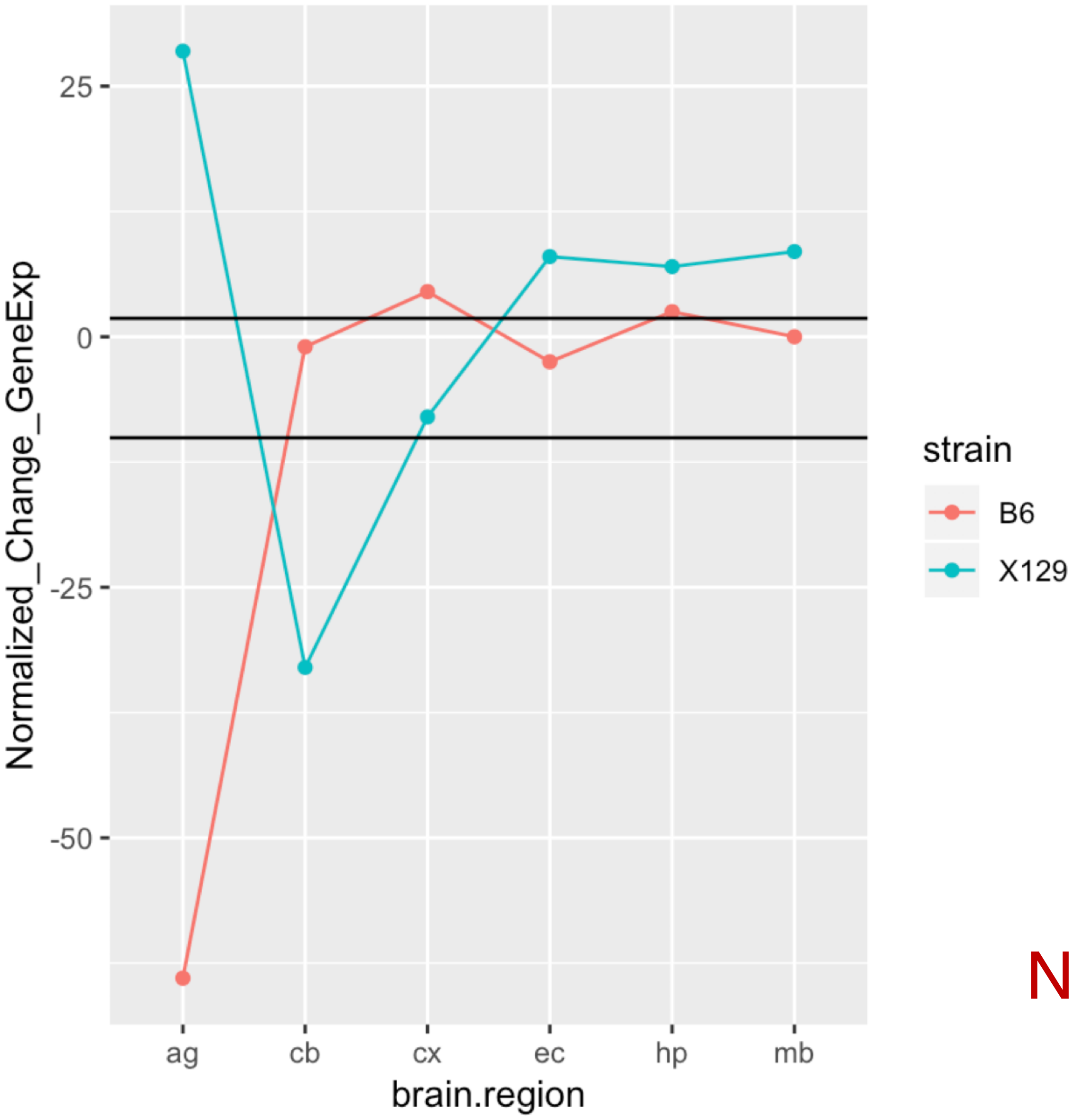
Case 3 - What are the significant effects?
a gene for which only the interaction is significant

gene aa051500.at - Only the interaction between brain regions and strain is significant (i.e., differences in mean gene expression levels of brain regions depend on strain, or vice-versa)



```
Response: aa051500.at
                    Df Sum Sq Mean Sq F value   Pr(>F)
strain               1  852.0  852.04  2.4399 0.144256
brain.region         5 2212.9  442.58  1.2674 0.339231
strain:brain.region  5 9087.2 1817.44  5.2045 0.009038 **
Residuals           12 4190.5  349.21
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

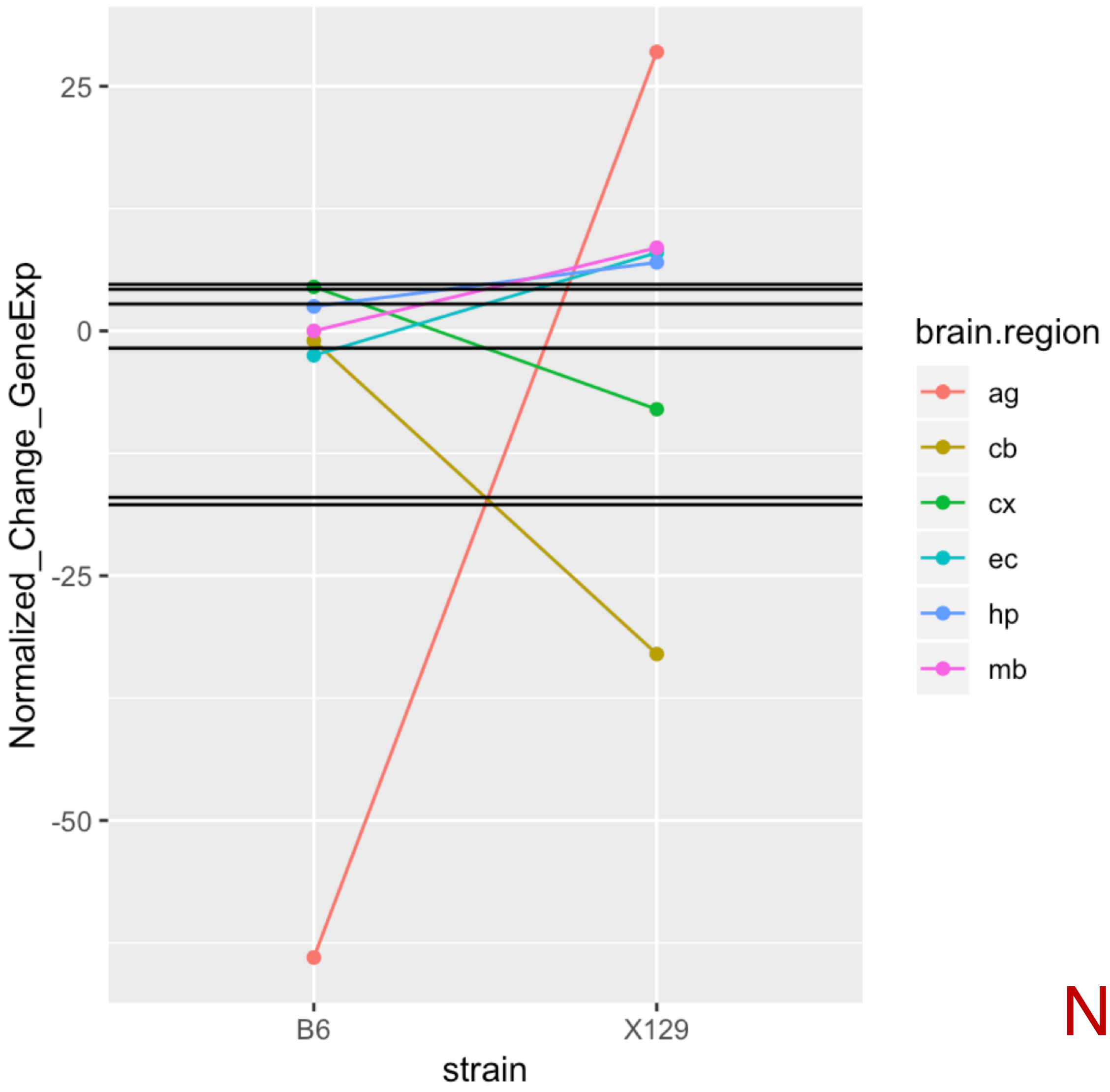
aa051500.at - Only the interaction between strain and brain region is significant



Black lines represent means of each strain independent of brain region and their difference is not significant – but some genes are more different than the other depending on the brain region

Note that the lines for strain are NOT parallel

aa051500.at - Only the interaction between strain and brain region is significant

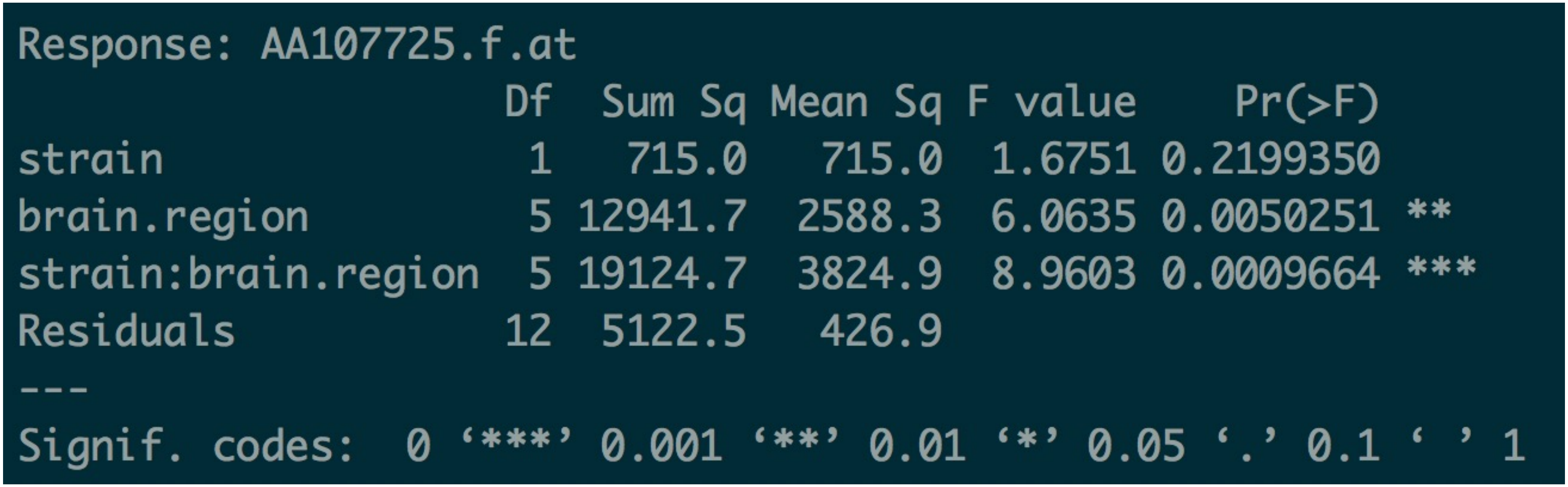


Black lines represent means of each brain region independent of strain and their differences are not significant – but some genes are more different than the other depending on the strain. There are 66 possible pairwise comparisons (tests).

Note that the lines for brain region are NOT parallel

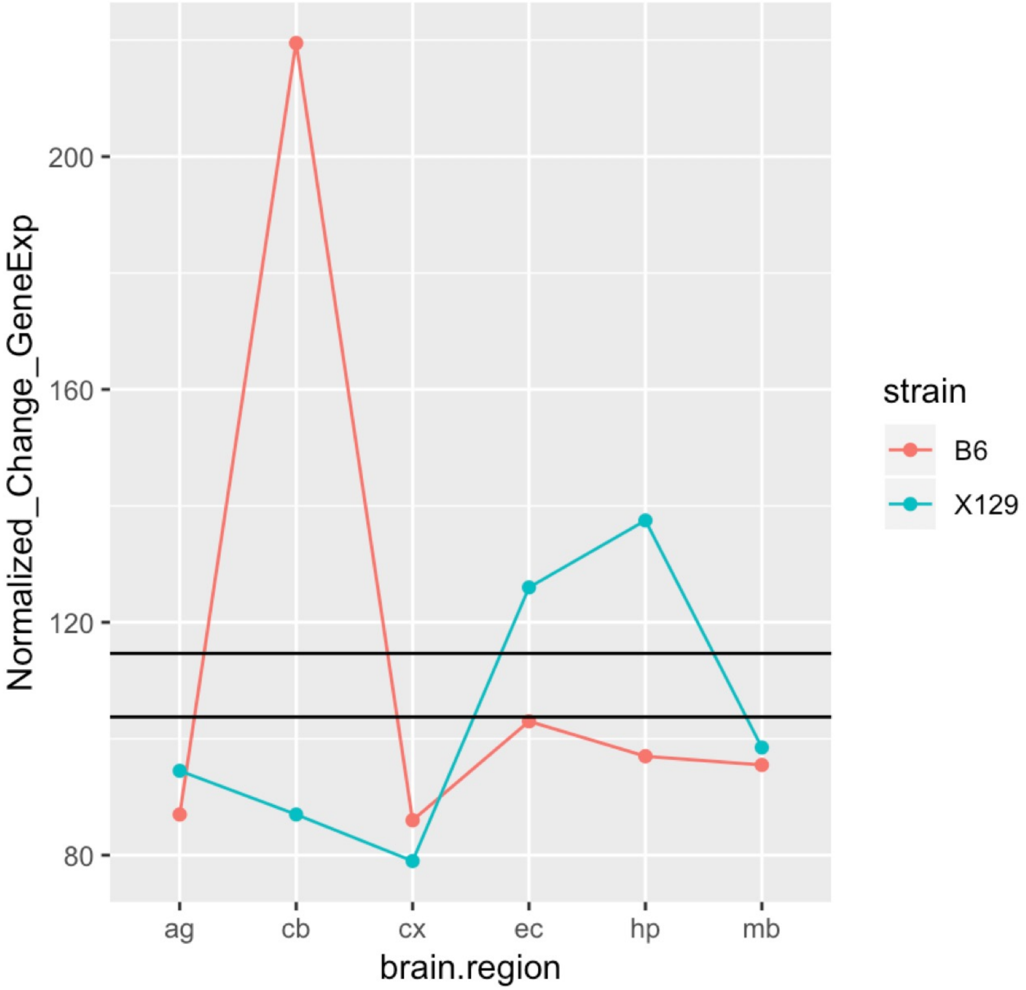
Case 4 - What are the significant effects? a gene for which at least one main factor and the interaction is significant

gene AA107725.f.at - The mean gene expression levels in brain regions vary, and the mean differences depend on the strain



```
Response: AA107725.f.at
                    Df  Sum Sq Mean Sq F value    Pr(>F)
strain               1   715.0   715.0  1.6751 0.2199350
brain.region         5 12941.7  2588.3  6.0635 0.0050251 **
strain:brain.region  5 19124.7  3824.9  8.9603 0.0009664 ***
Residuals           12  5122.5   426.9
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

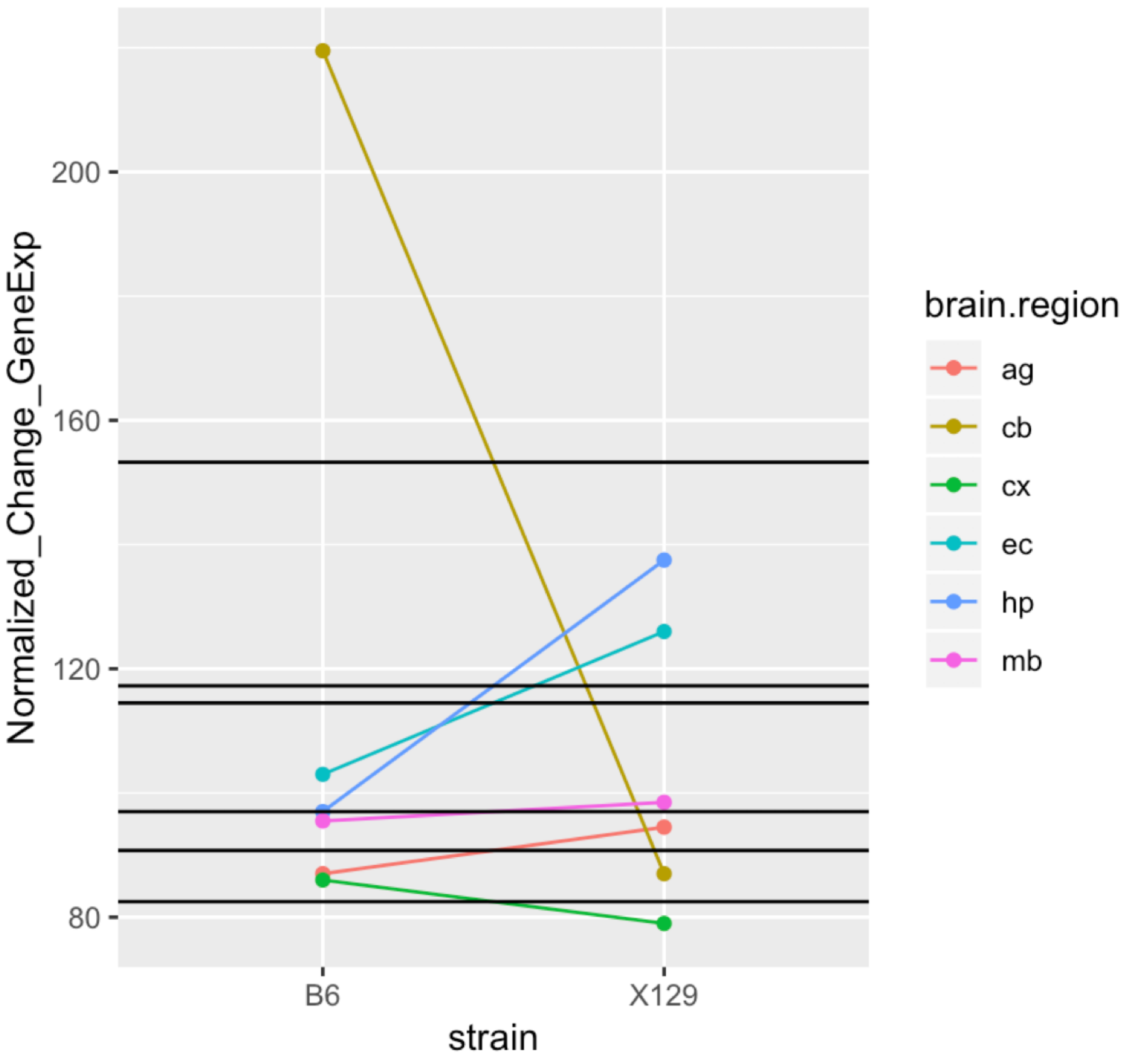
AA107725.f.at – brain region differs in gene expression and interaction is significant (only interpret interactions seem to be interpretable)



Black lines represent means of each strain independent of brain region and their differences are not significant – but some genes are more different than the other depending on the brain region

Note that the lines are NOT parallel

AA107725.f.at – brain region differs in gene expression and interaction is significant (only interpret interactions seem to be interpretable)



Black lines represent means of each brain region independent of strain and their differences are significant – and some genes are more different than the other depending on the strain

Note that the lines are NOT parallel

A word on balanced designs

The ANOVAs performed here (and in tutorial 3) are based on equal number of observations per combination of groups.

In the fictional diet example, there are 5 individuals in each of the 4 combinations of diet (yes/no) and exercise (yes/no).

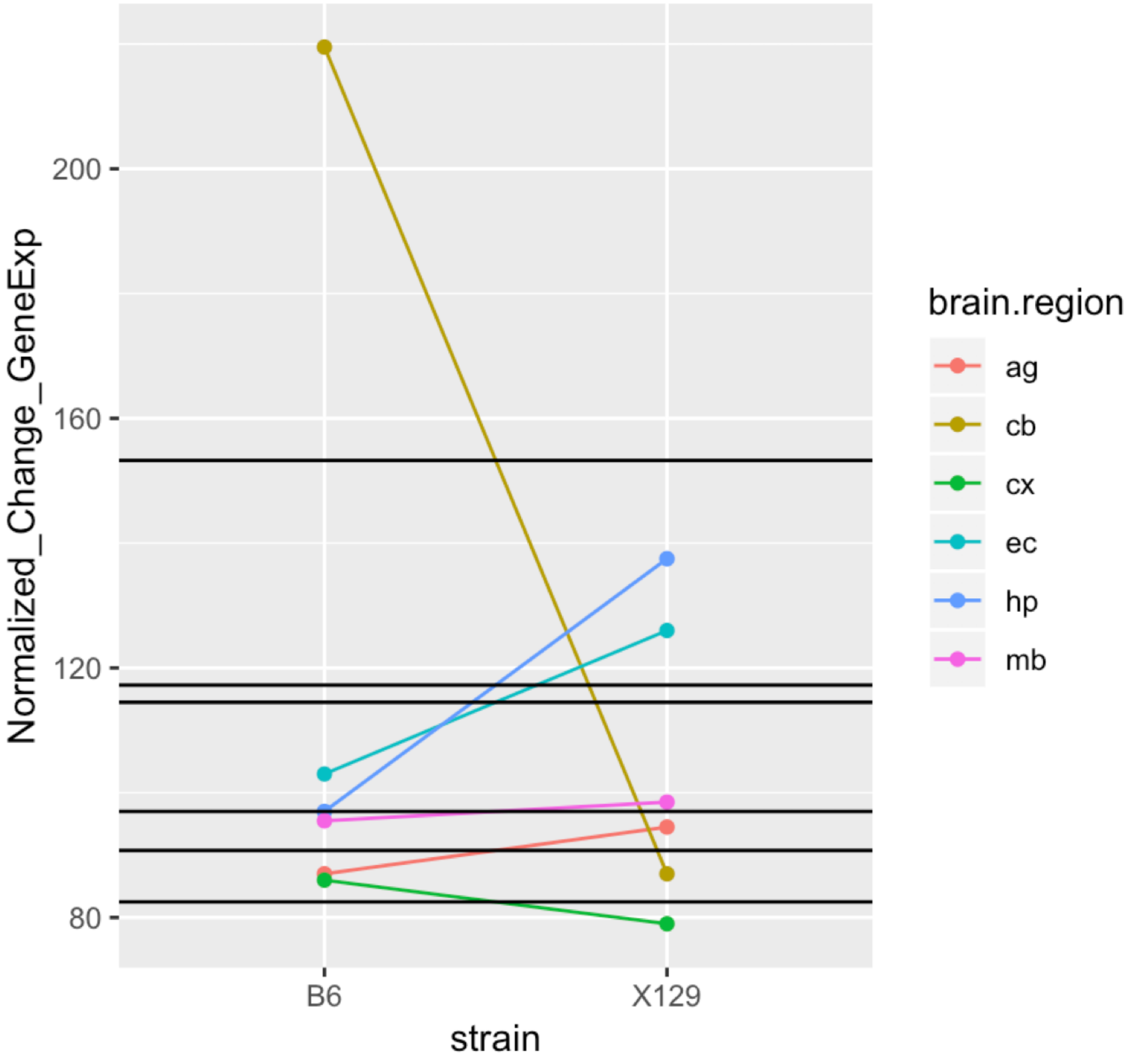
In the gene expression study, there are 2 individuals in each of the 12 combinations of strain (2 strains) and brain region (6 regions).

For balanced designs, we say that the design is fully orthogonal because there is no variation that is shared between factors (a concept we will see in a few lectures; under ANCOVA).

For fully orthogonal designs, we can use what is called a Type I Sum-of-Squares (Type I SS) or a Type III SS. When factors are not fully orthogonal, then we use the Type III SS (Sum-of-Squares). We will learn about Type III in the ANCOVA module).

Lecture 6: Which effects are significant?

which pairwise means to compare?


Normalized_Change_GeneExp
B6
X129
strain
brain.region
ag
cb
cx
ec
hp
mb