Stat...

QUEST!!!

Stat...

QUEST!!!

Stat...

Stat...

QUEST!!!

Stat...

QUEST!!!

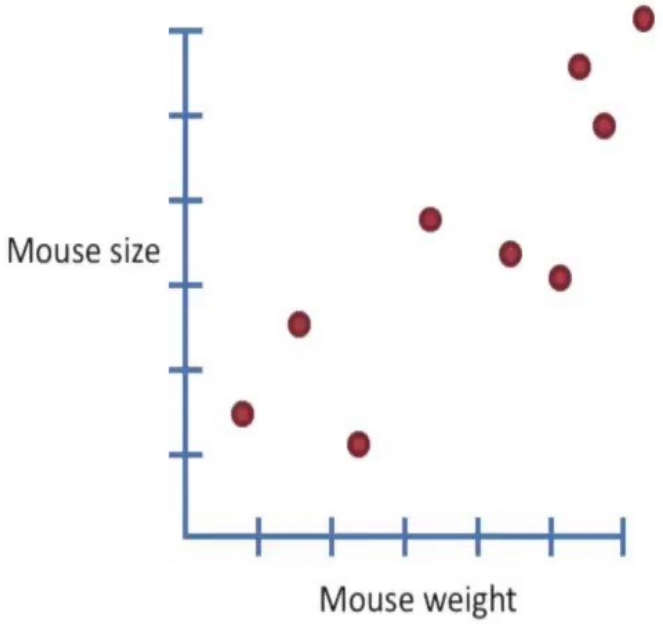
Stat...

QUEST!!!

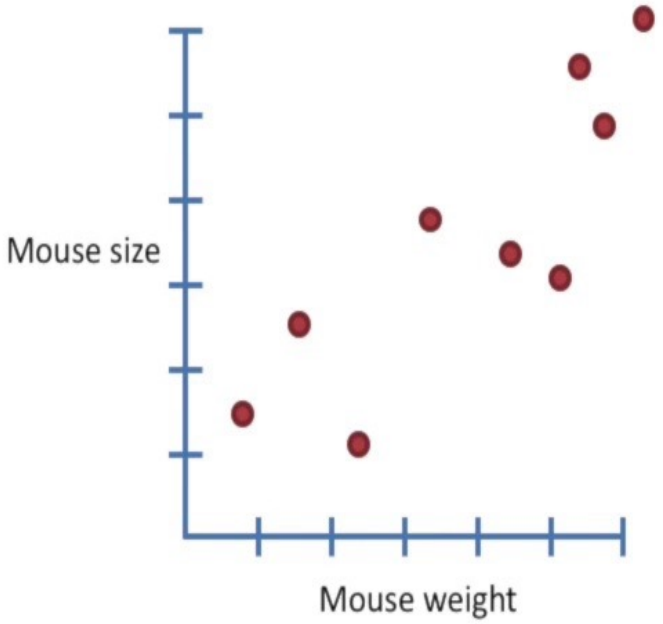
Yeah!!!

StatQuest: General Linear Models
Part 2:
t-tests and ANOVA

Quick Review of Linear Regression…

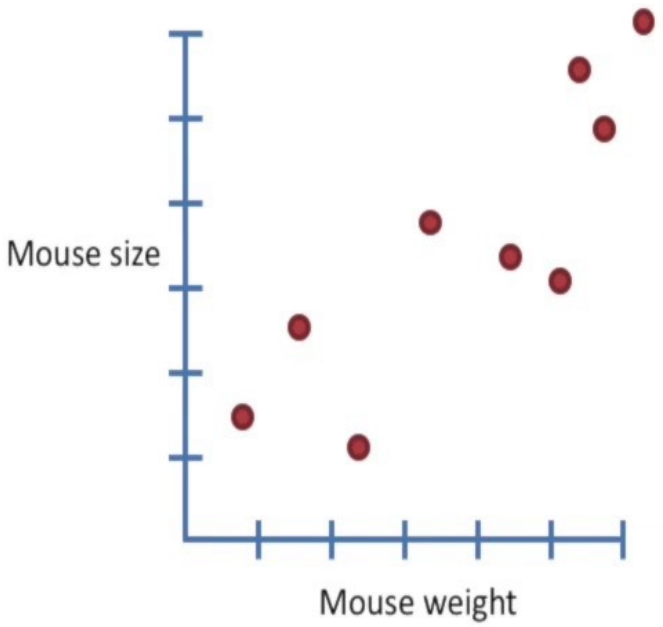


We measured **mouse weight** and **mouse size** and we wanted learn 2 things from it:



We measured **mouse weight** and **mouse size** and we wanted learn 2 things from it:

1) How useful was mouse weight for predicting mouse size? (R^2 told us this).

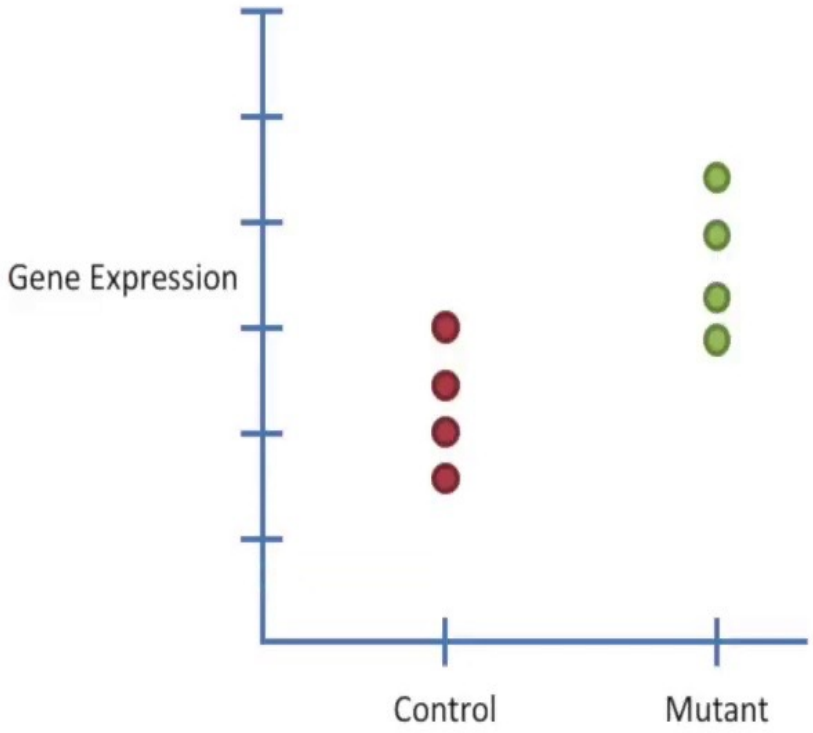


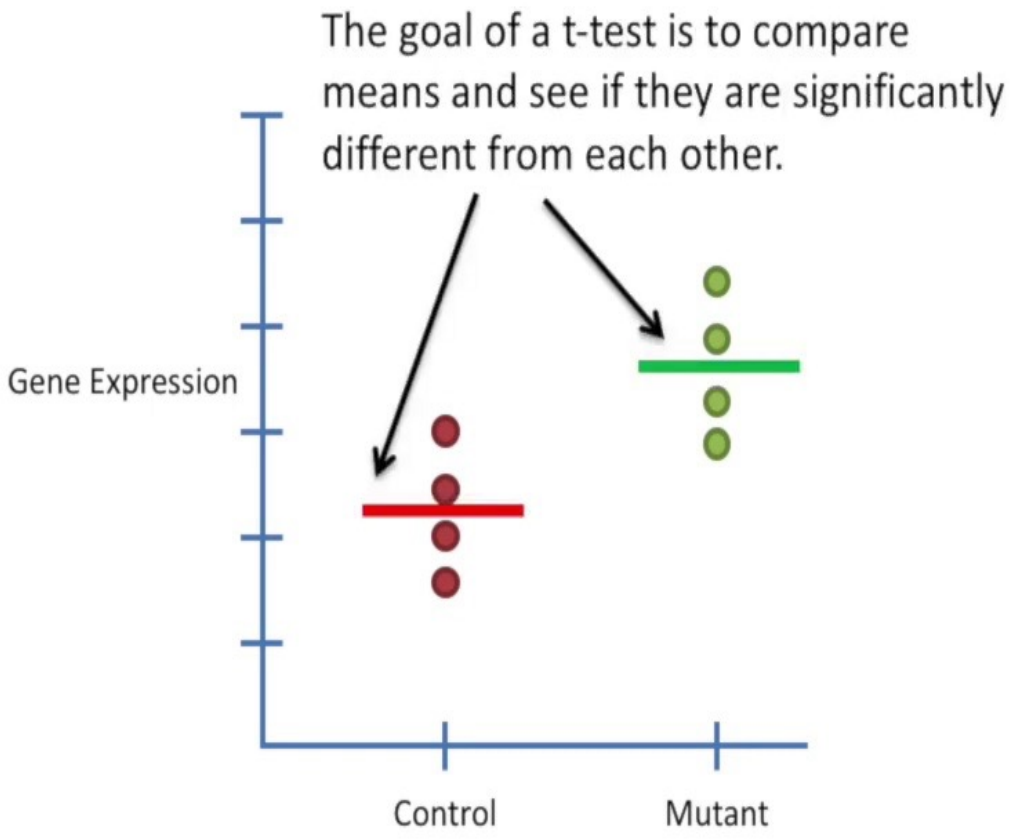
We measured **mouse weight** and **mouse size** and we wanted learn 2 things from it:

1) How useful was mouse weight for predicting mouse size? (R^2 told us this).

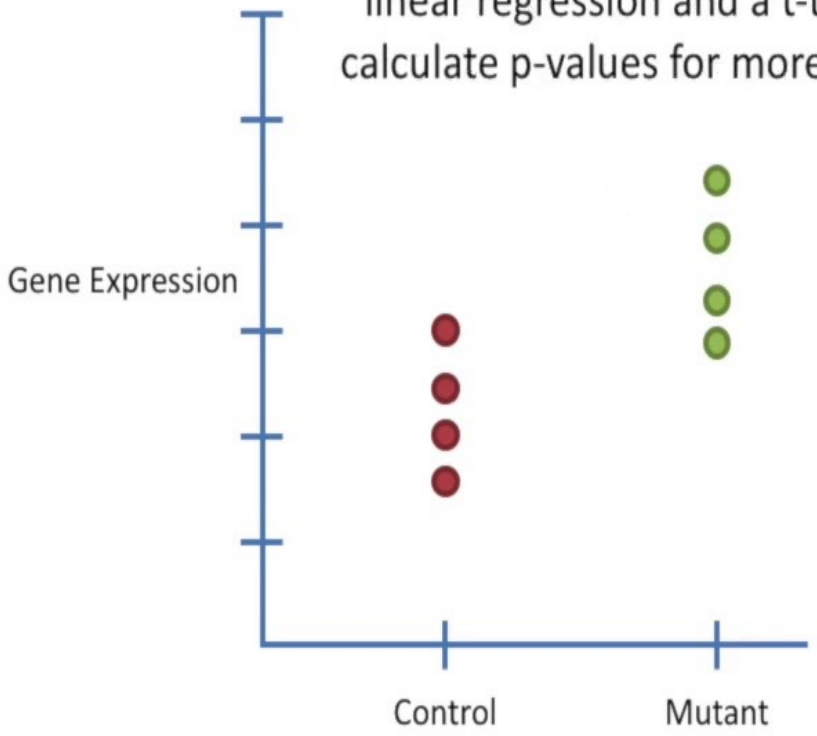
2) Was that relationship due to chance? (The ***p*-value** told us this.)

Now let's see if we can apply those concepts to a t-test.

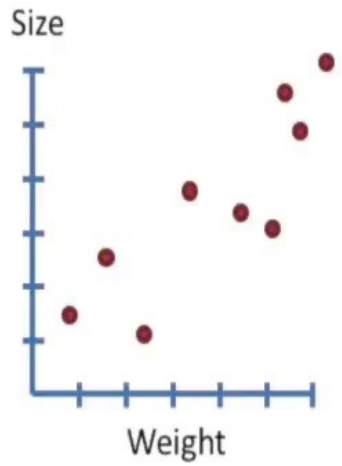



The goal of a t-test is to compare means and see if they are significantly different from each other.
Gene Expression
Control
Mutant

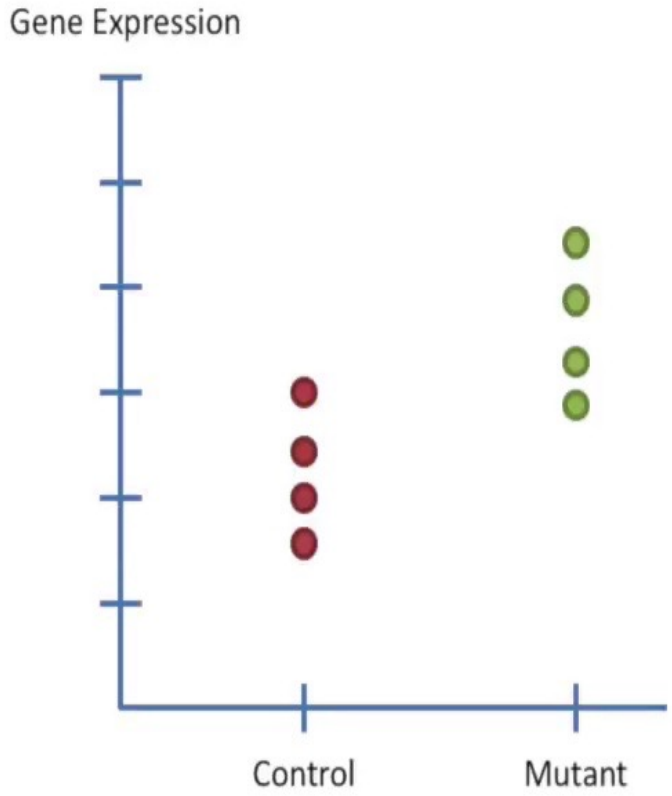
If the same method can calculate *p*-values for a linear regression and a t-test, then we can easily calculate p-values for more complicated situations.



Linear Regression



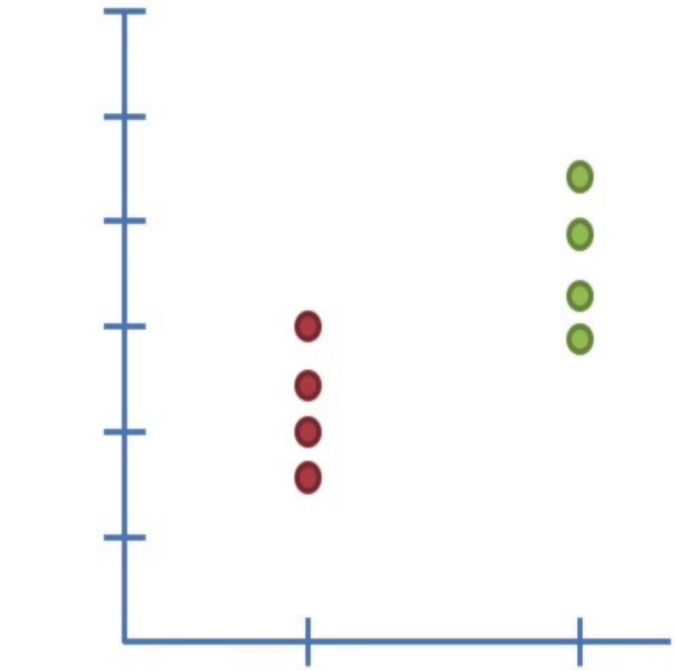
t-test



Step 1: Ignore the x-axis and find the overall mean.

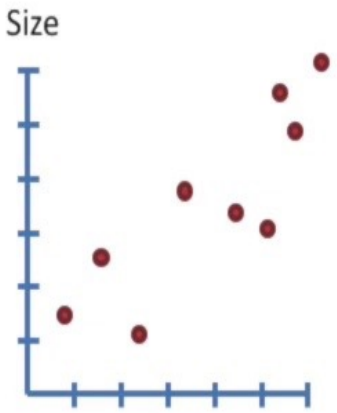
t-test

Gene Expression



Linear Regression

Size



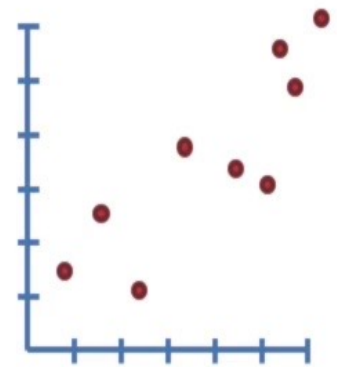
Step 1: Ignore the x-axis and find the overall mean.

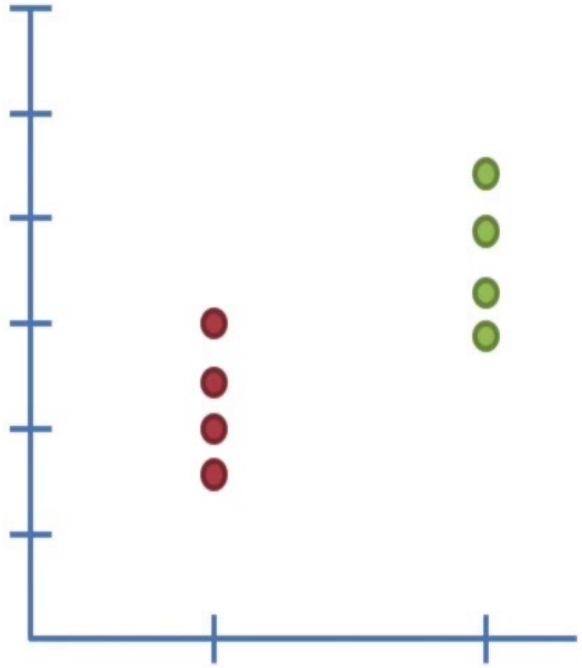
t-test

Gene Expression

Linear Regression

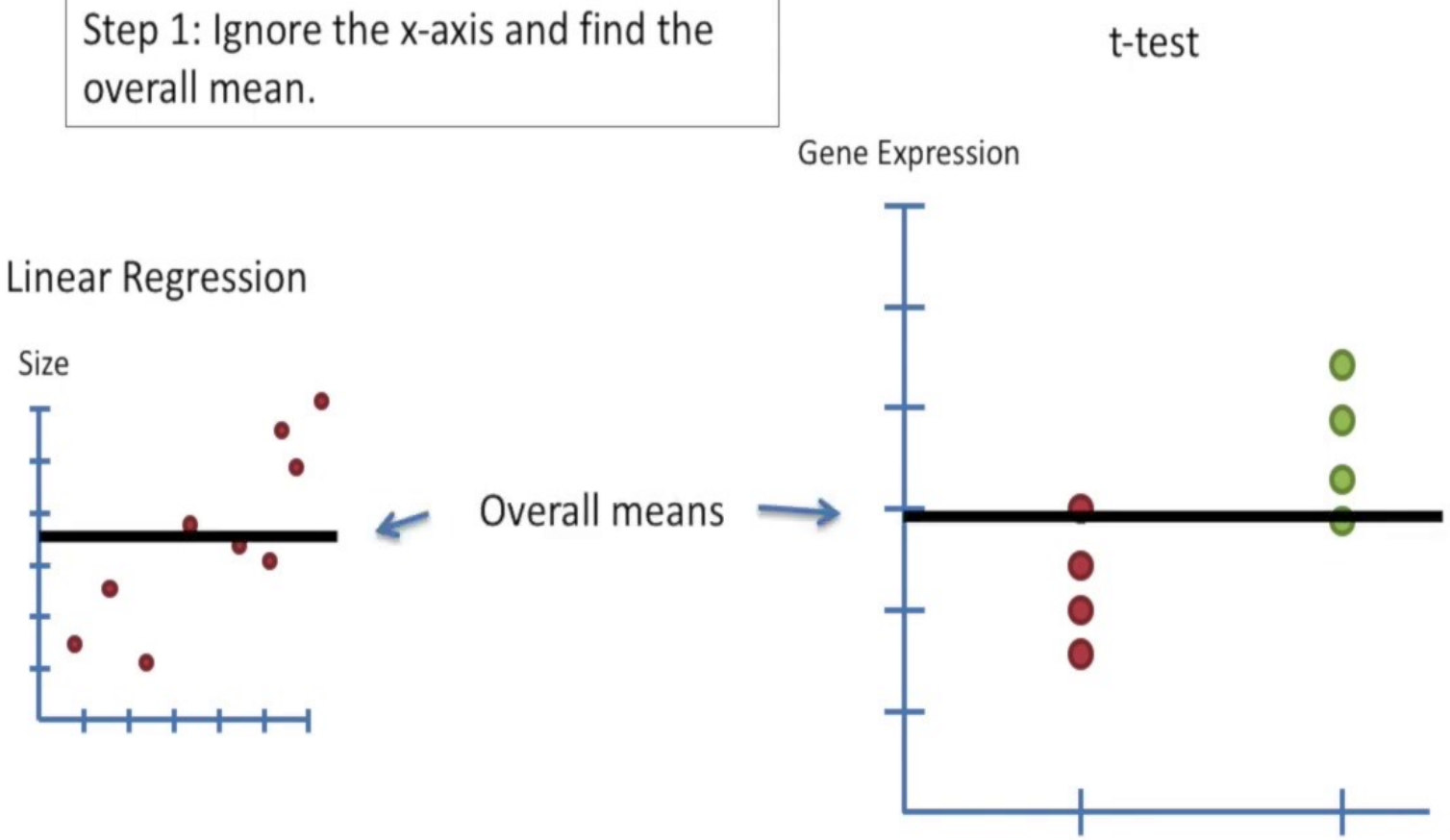
Size





To emphasize that we want to focus on the y-axis, I've removed the labels on the x-axis.

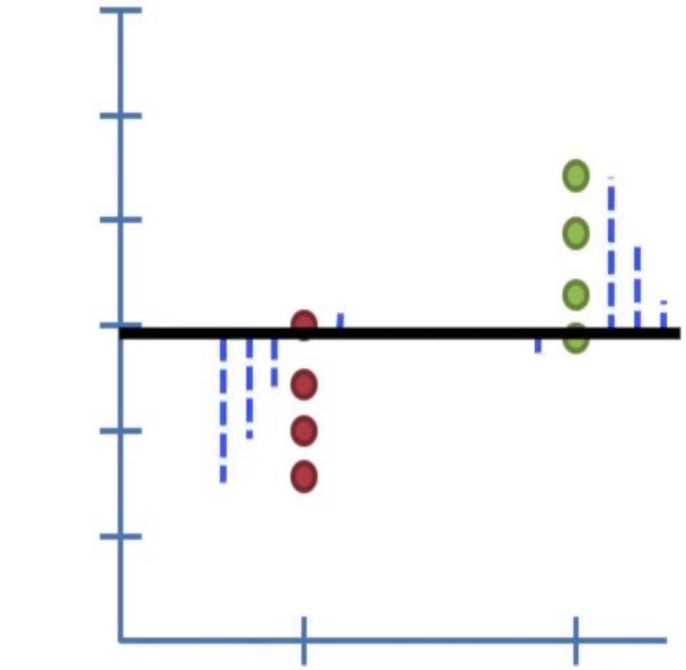
Step 1: Ignore the x-axis and find the overall mean.



Step 2: Calculate SS(mean), the sum of squared residuals around the mean.

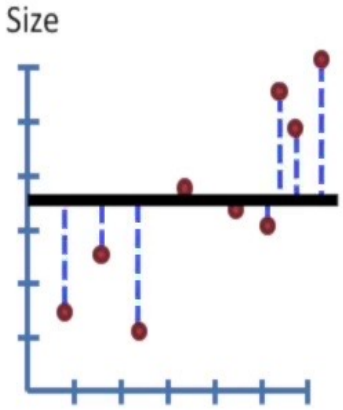
t-test

Gene Expression



Linear Regression

Size



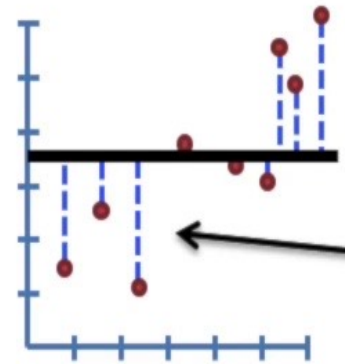
Step 2: Calculate SS(mean), the sum of squared residuals around the mean.

t-test

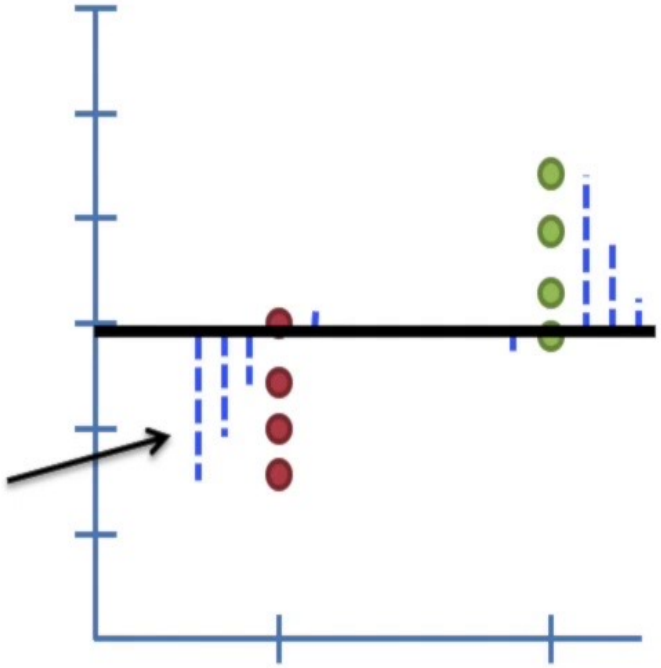
Gene Expression

Linear Regression

Size



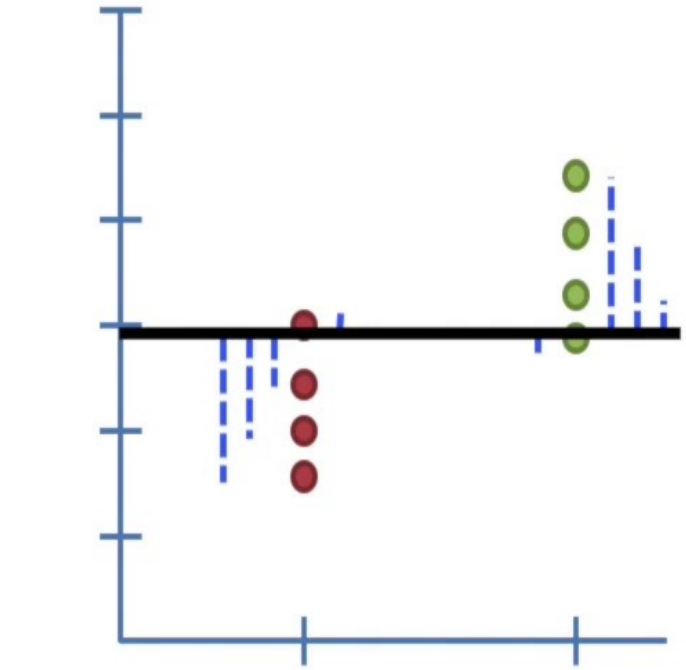
These are the residuals, the distance from the data points to the lines.



Step 2: Calculate SS(mean), the sum of squared residuals around the mean.

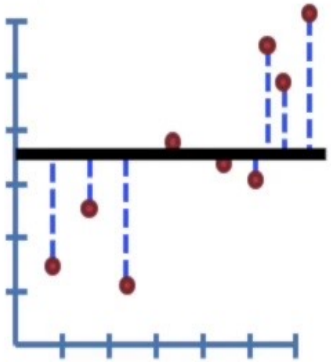
t-test

Gene Expression



Linear Regression

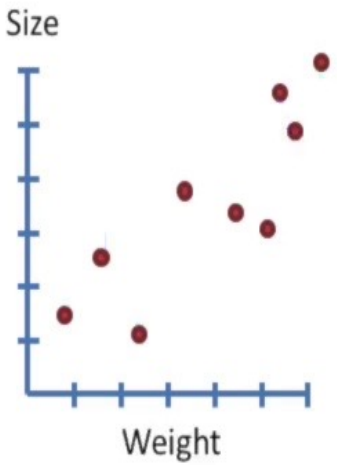
Size



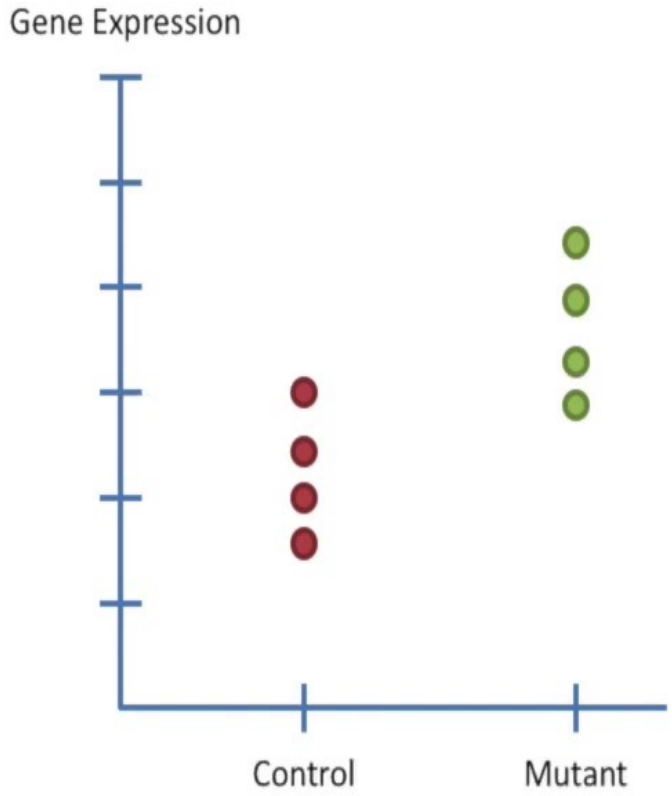
BAM! Calculating SS(mean) was easy!

Step 3: Fit a line to the data.

Linear Regression



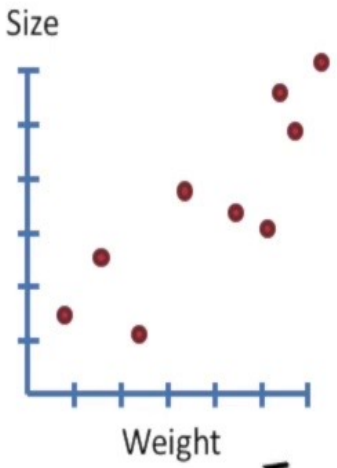
t-test

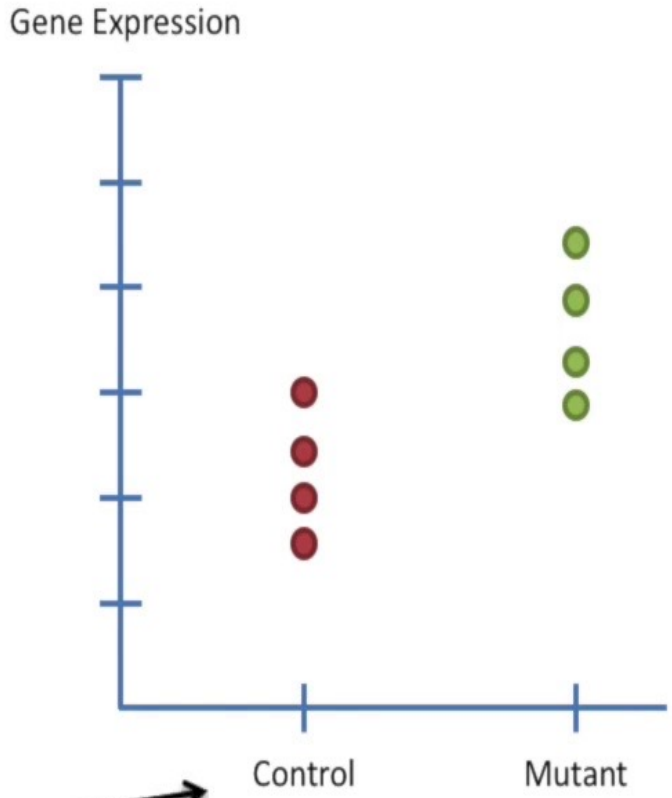


Step 3: Fit a line to the data.

t-test

Linear Regression

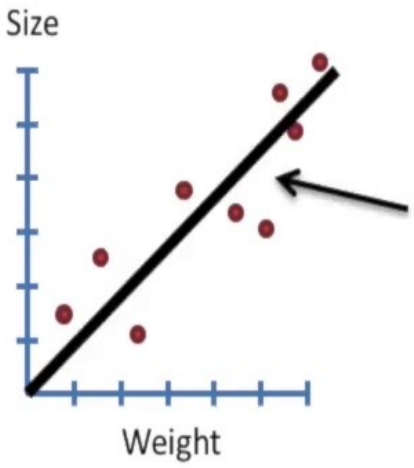




NOTE: This is when we start caring about the x-axis again...

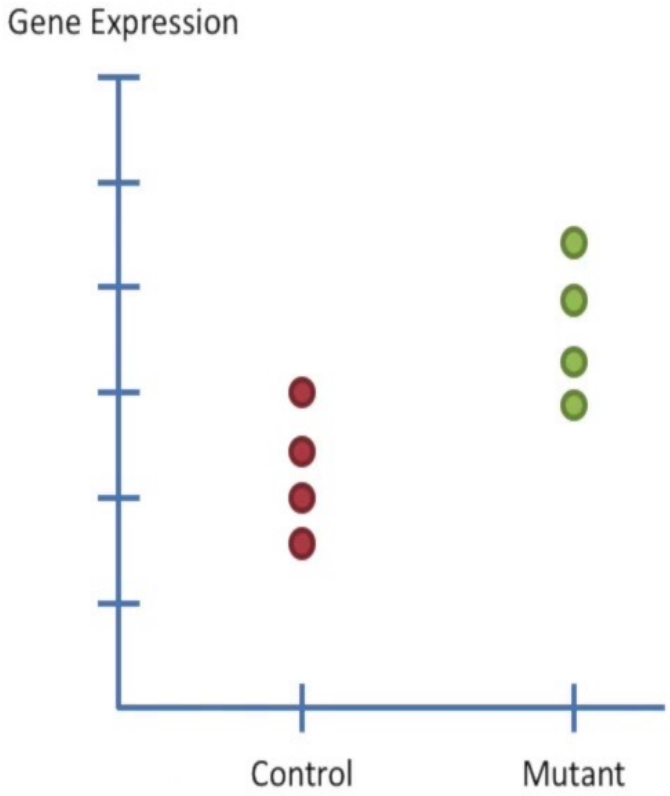
Step 3: Fit a line to the data.

Linear Regression



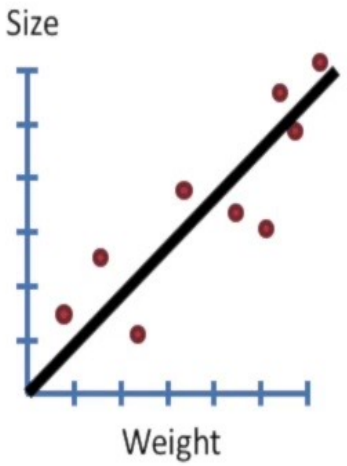
The least-squares fit to the data.

t-test

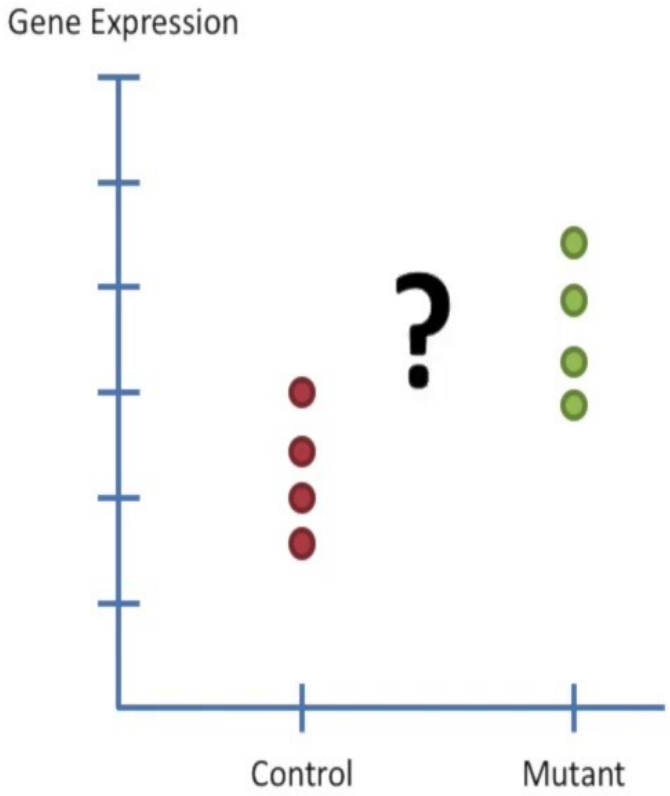


Step 3: Fit a line to the data.

Linear Regression



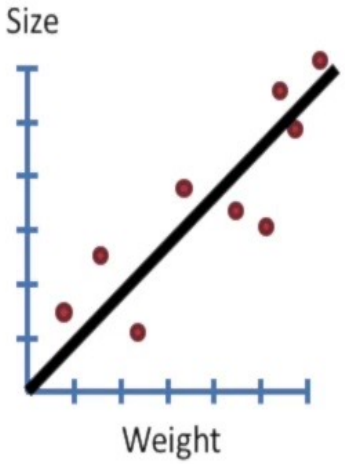
t-test

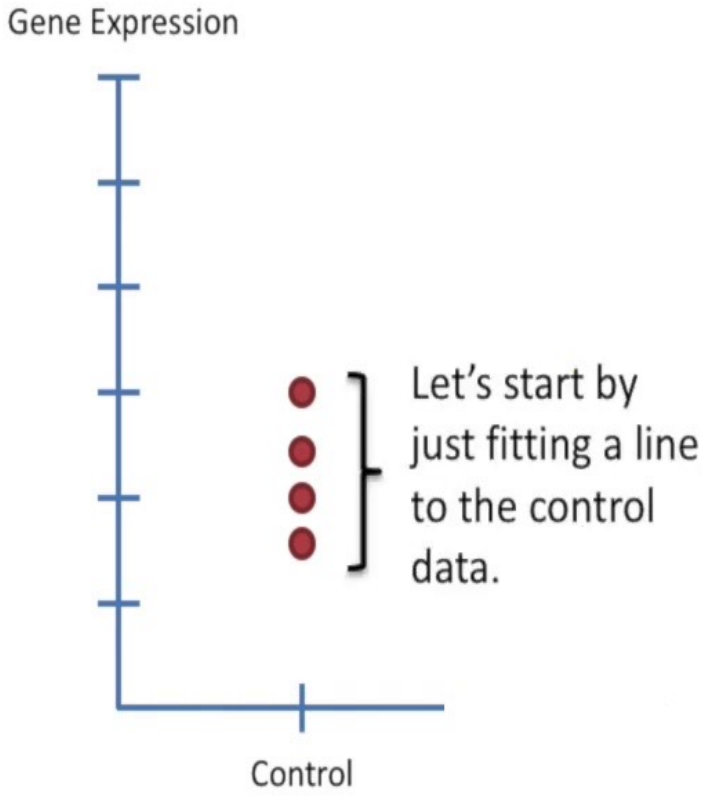


Step 3: Fit a line to the data.

t-test

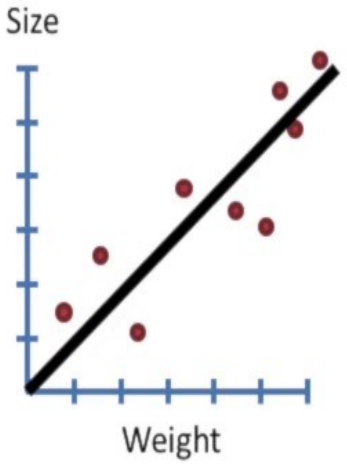
Linear Regression



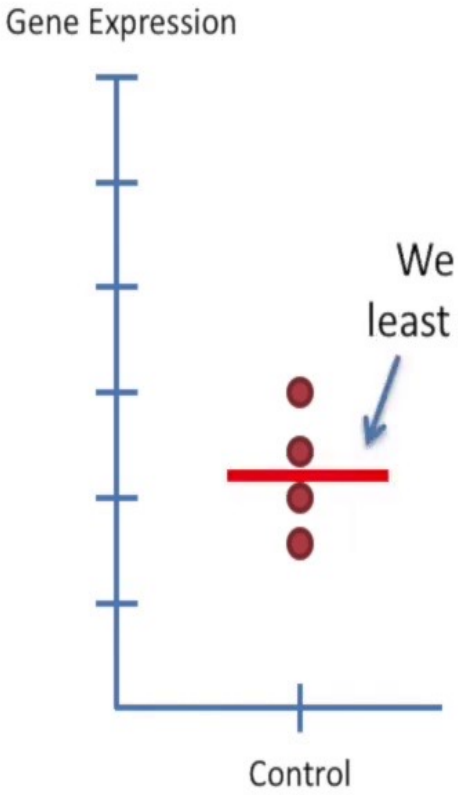


Step 3: Fit a line to the data.

Linear Regression



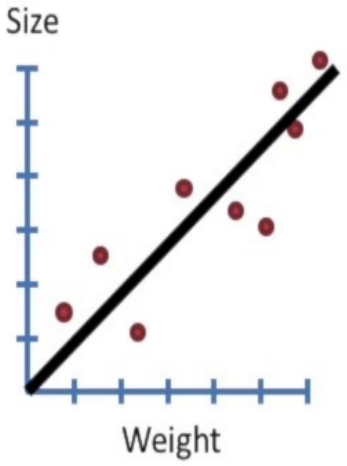
t-test



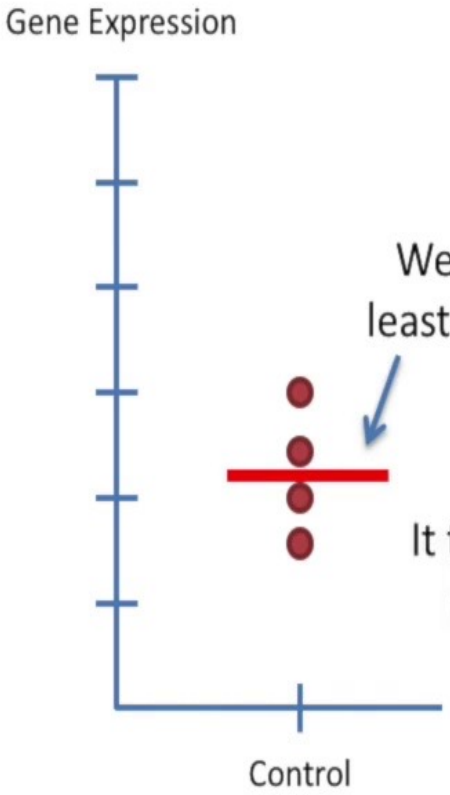
We start by finding a least squares fit to the control data.

Step 3: Fit a line to the data.

Linear Regression



t-test

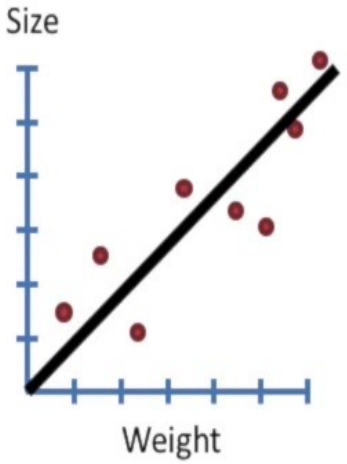


We start by finding a least squares fit to the control data.

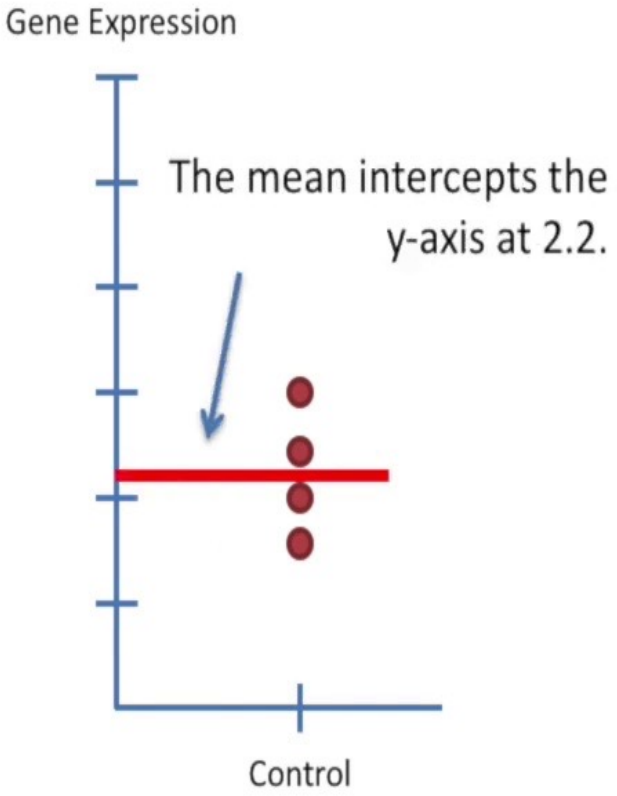
It turns out that the mean is the least-squares fit.

Step 3: Fit a line to the data.

Linear Regression

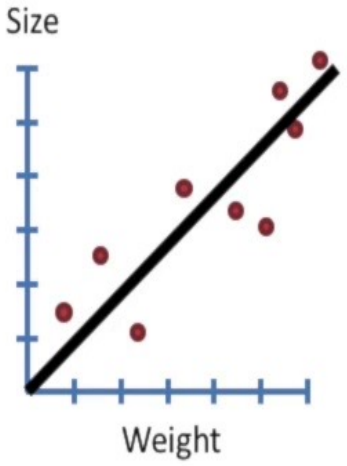


t-test

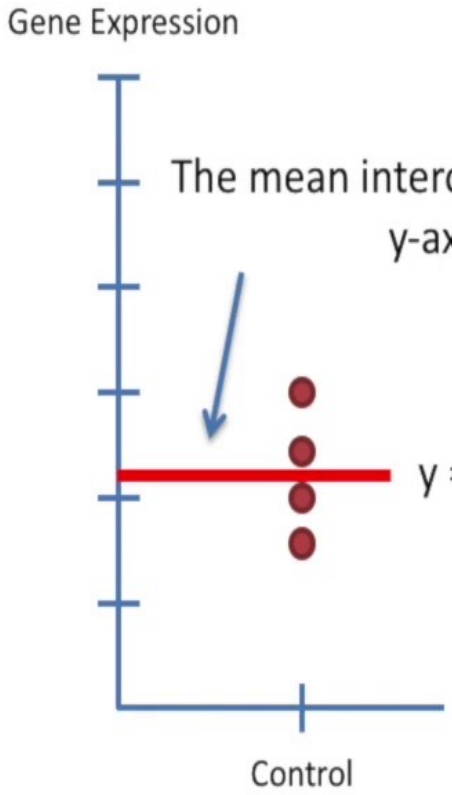


Step 3: Fit a line to the data.

Linear Regression



t-test



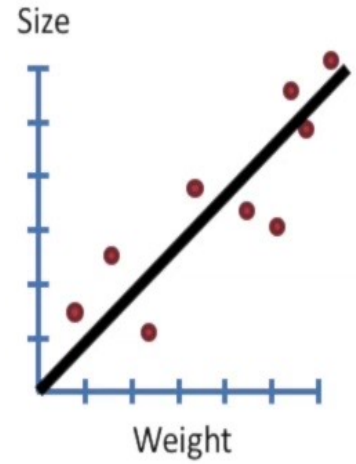
The mean intercepts the y-axis at 2.2.

$y = 2.2$

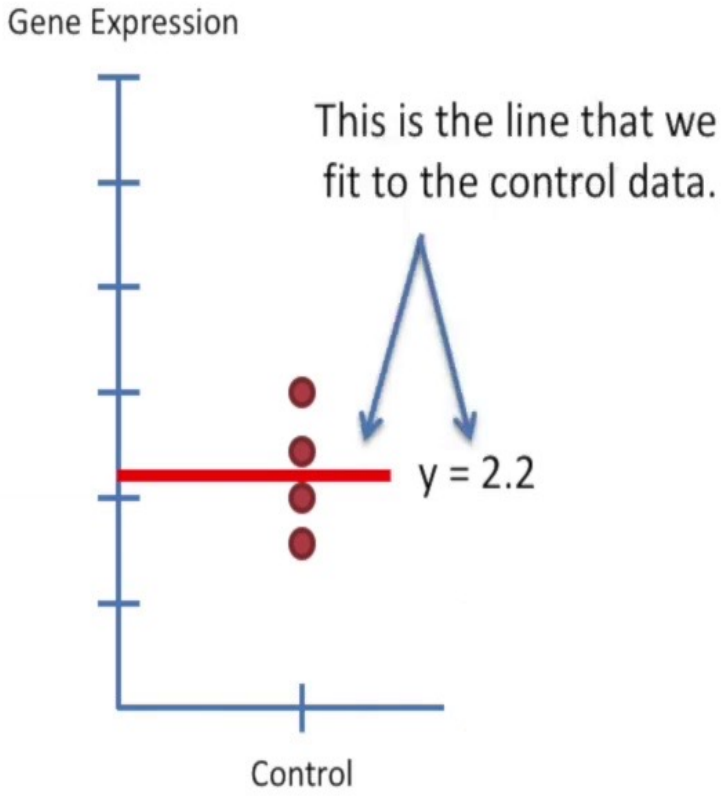
This is the equation for a horizontal line that intercepts the y-axis at 2.2.

Step 3: Fit a line to the data.

Linear Regression



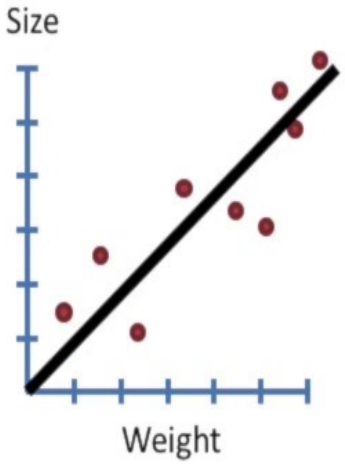
t-test

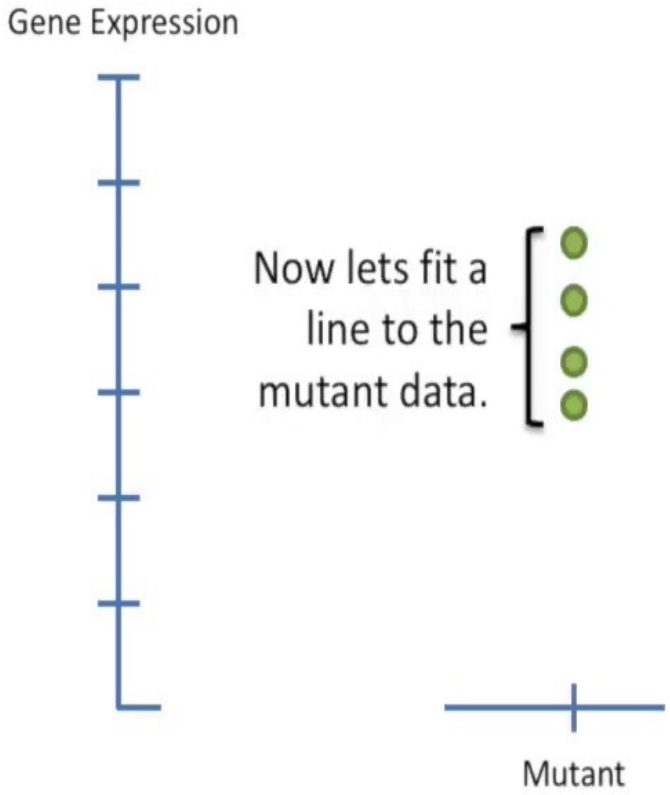


Step 3: Fit a line to the data.

t-test

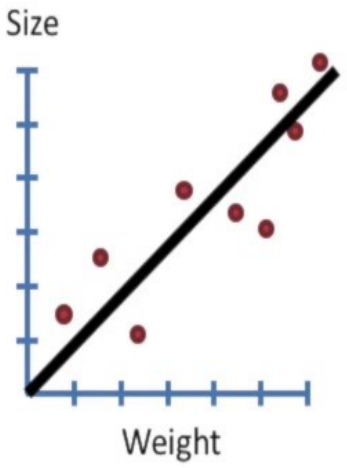
Linear Regression



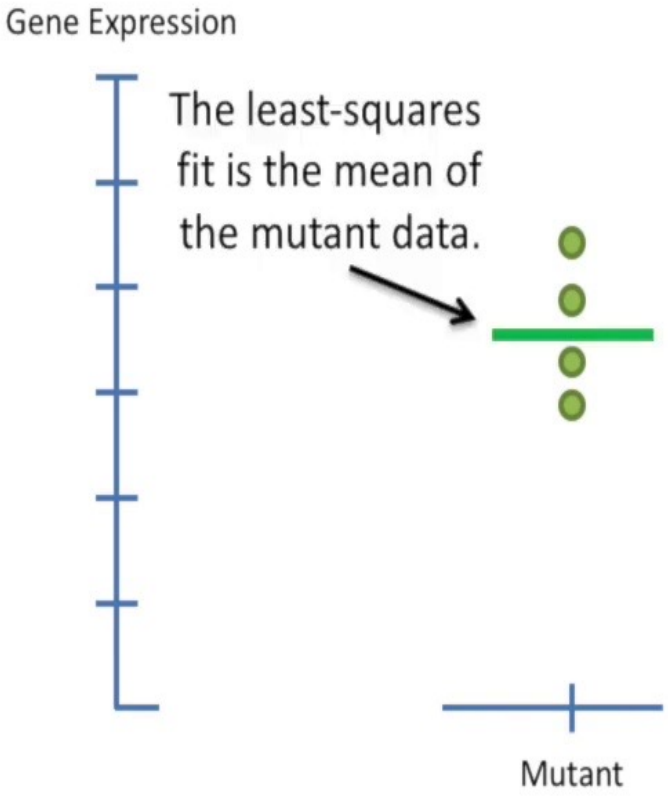


Step 3: Fit a line to the data.

Linear Regression

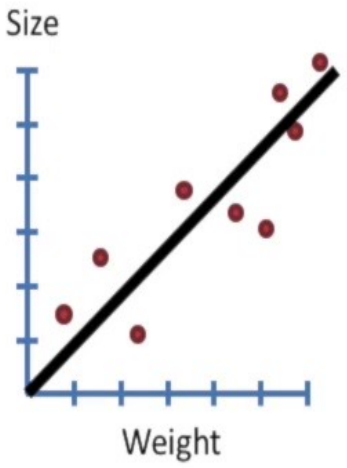


t-test

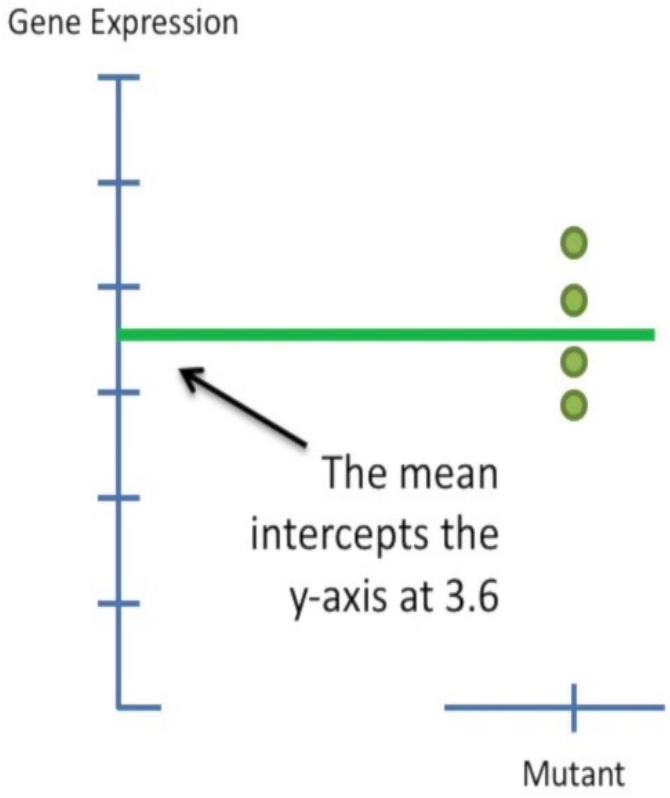


Step 3: Fit a line to the data.

Linear Regression

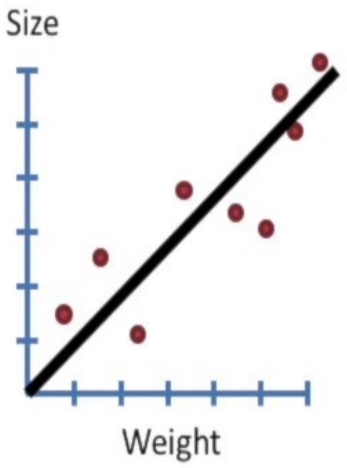


t-test

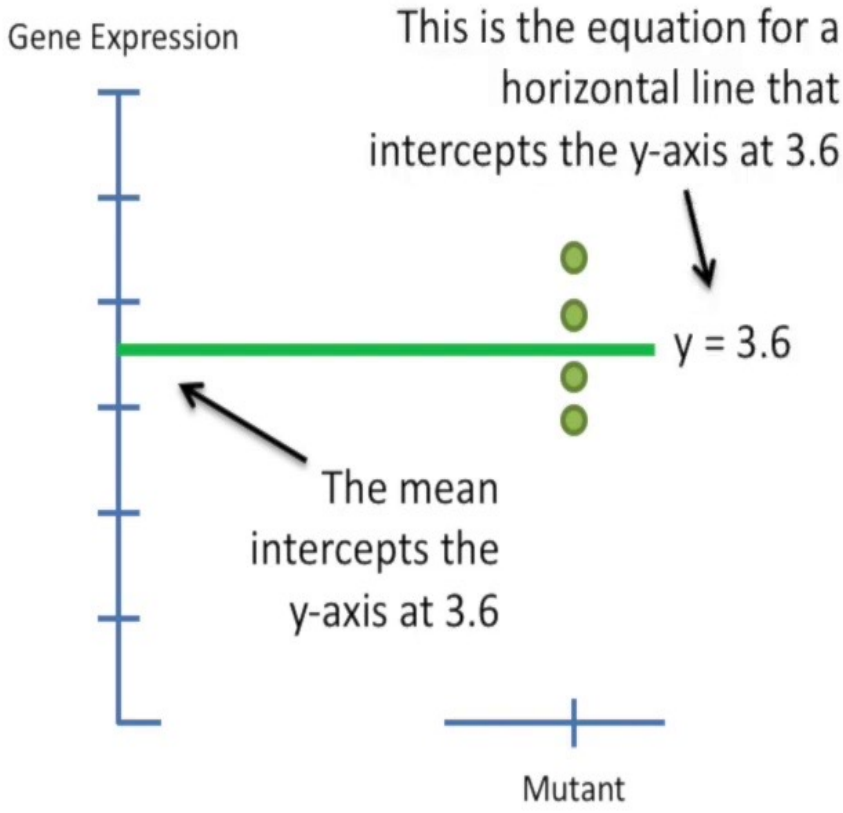


Step 3: Fit a line to the data.

Linear Regression

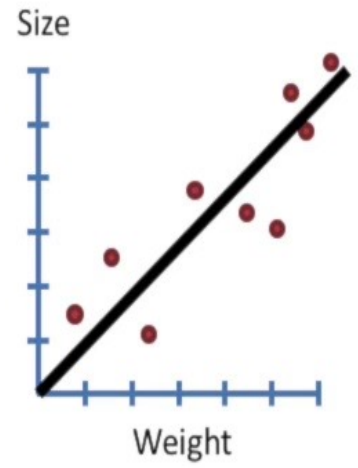


t-test

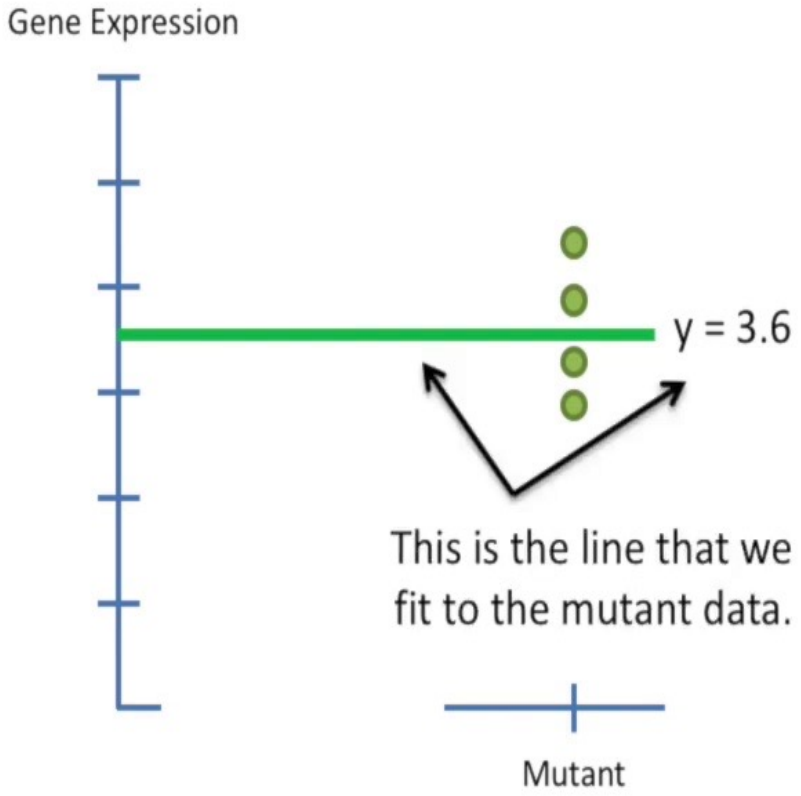


Step 3: Fit a line to the data.

Linear Regression



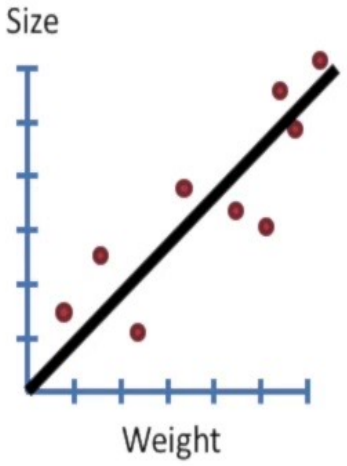
t-test

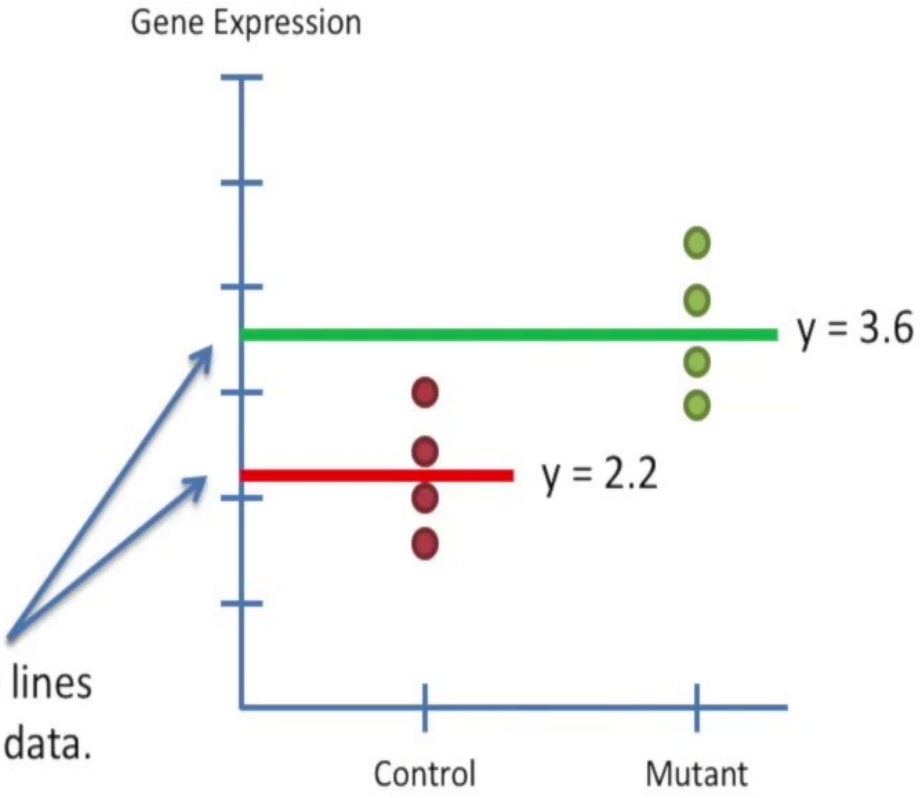


Step 3: Fit a line to the data.

t-test

Linear Regression



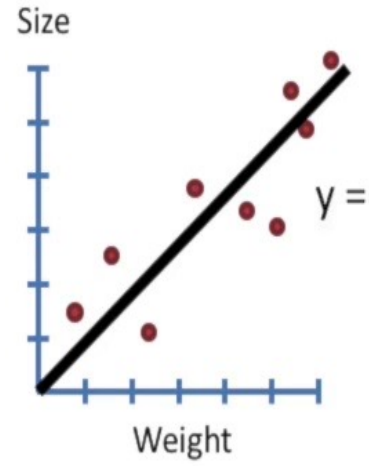


We have fit two lines to the data.

Step 3: Fit a line to the data.

t-test

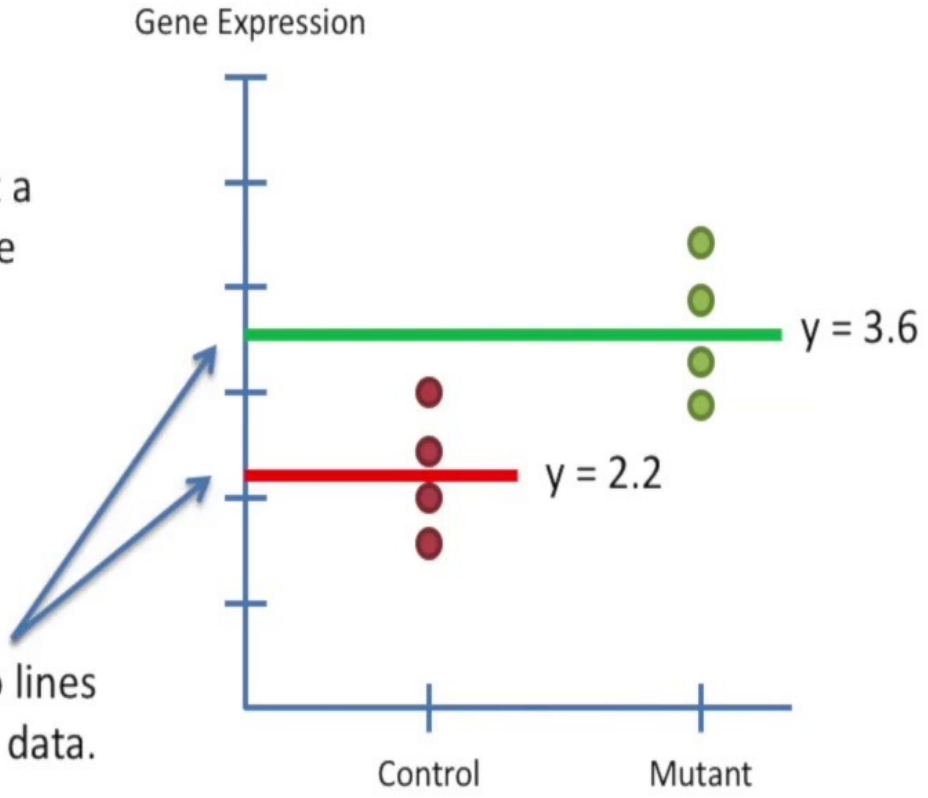
Linear Regression



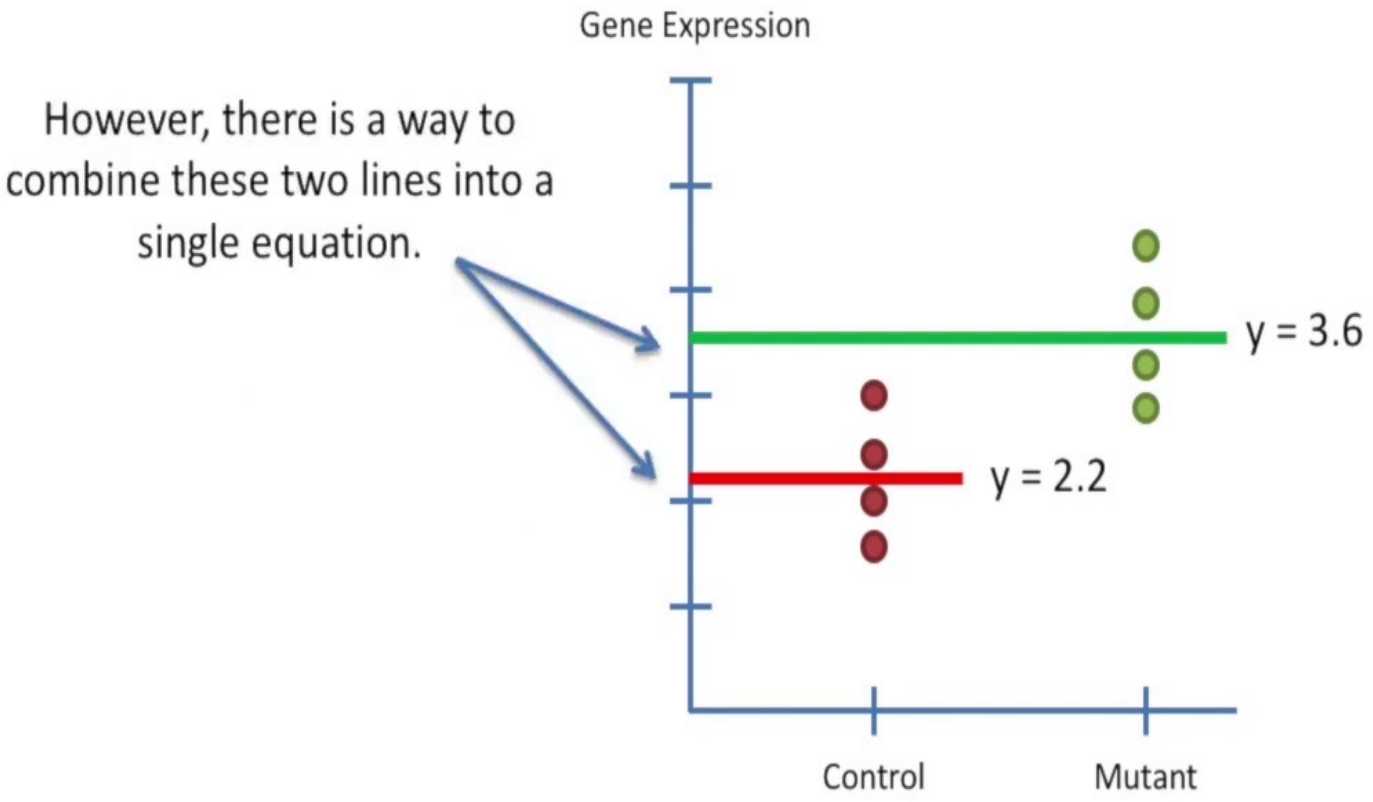
Originally we fit a single line to the data.

$y = 0.1 + 0.6 \times \text{weight}$

We have fit two lines to the data.



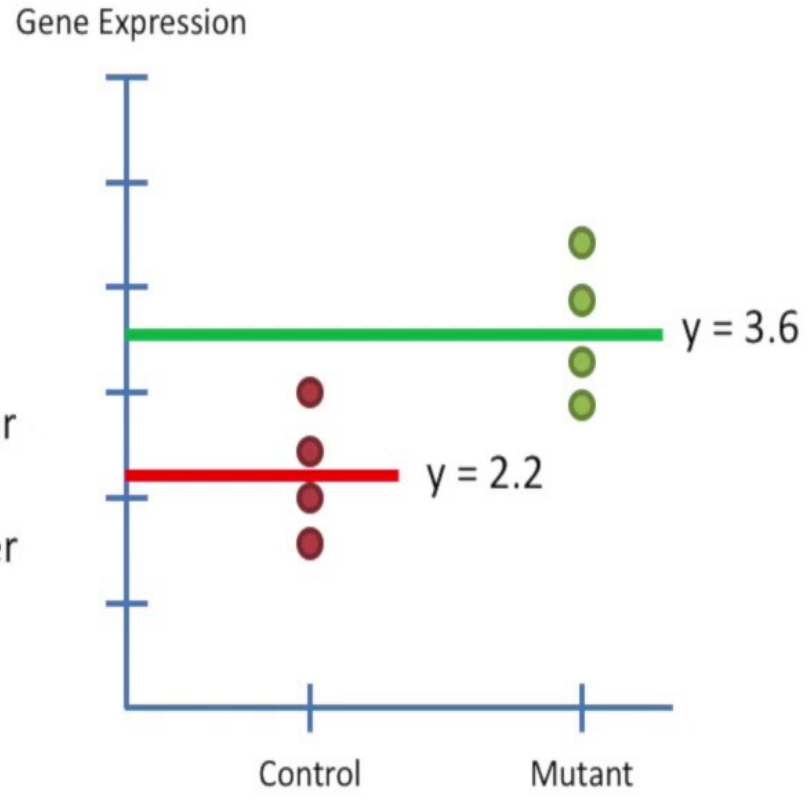
t-test


However, there is a way to combine these two lines into a single equation.
Gene Expression
y = 3.6
y = 2.2
Control
Mutant

t-test

However, there is a way to combine these two lines into a single equation.

This will make the steps for computing "*F*" the exact same for the regression and the t-test, which, in turn, means a computer can do it automatically.

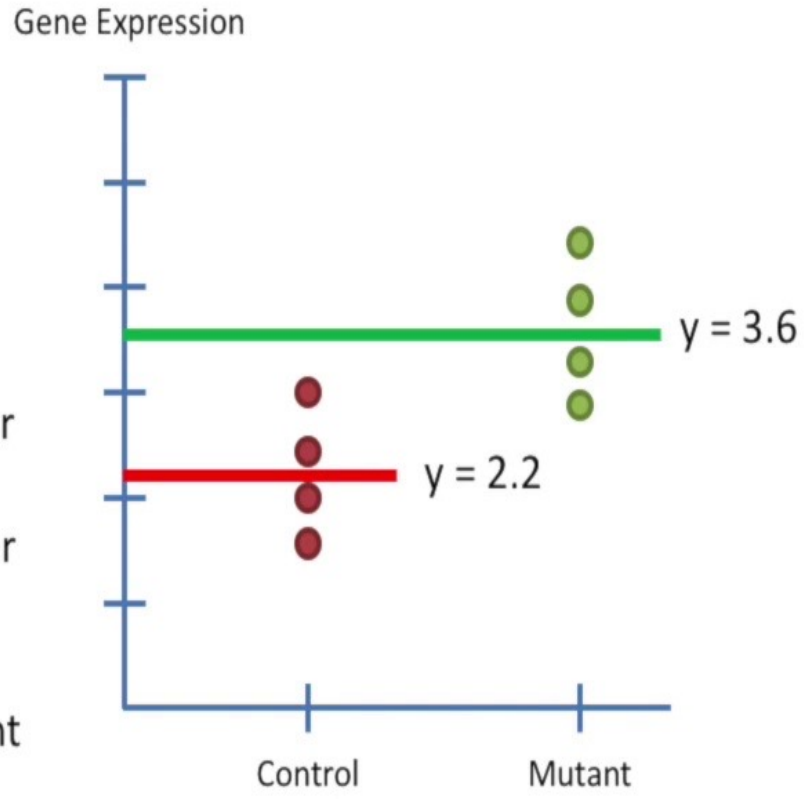


t-test

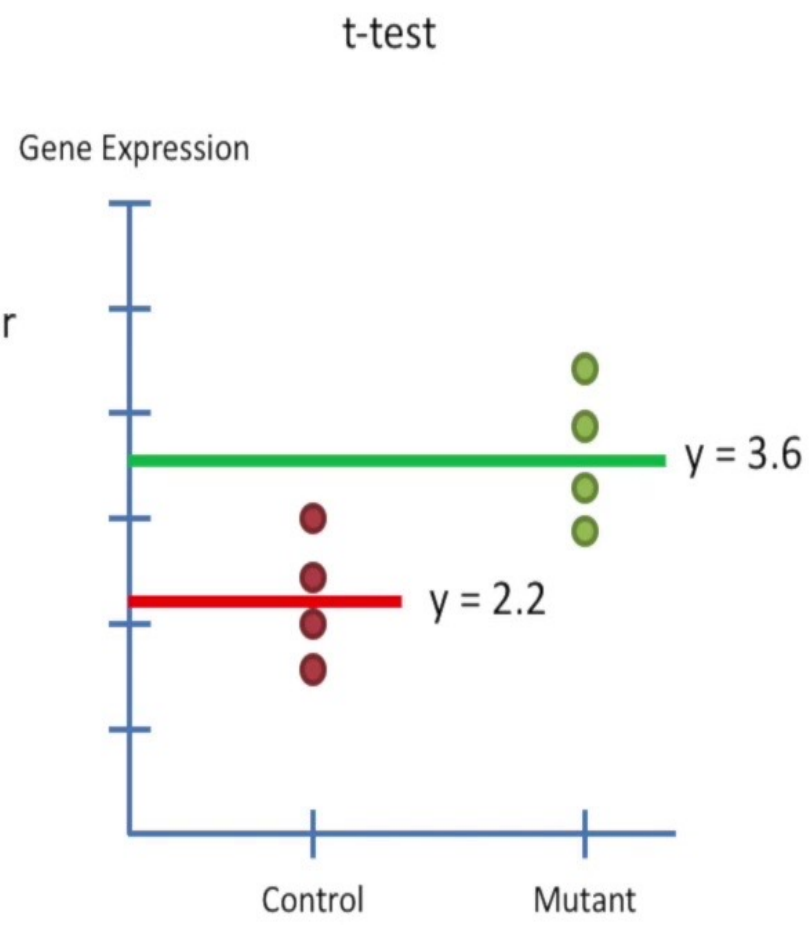
However, there is a way to combine these two lines into a single equation.

This will make the steps for computing "*F*" the exact same for the regression and the t-test, which, in turn, means a computer can do it automatically.

This is key, because we don't want to do this by hand, ever....

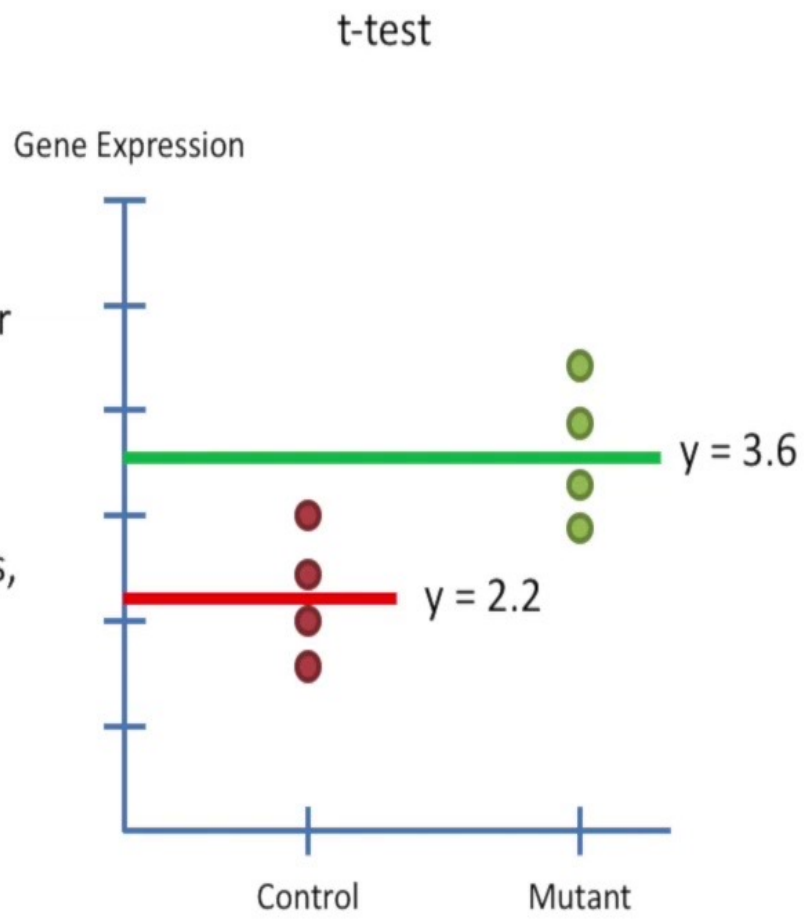


This is going to look weird, but just bear with me.



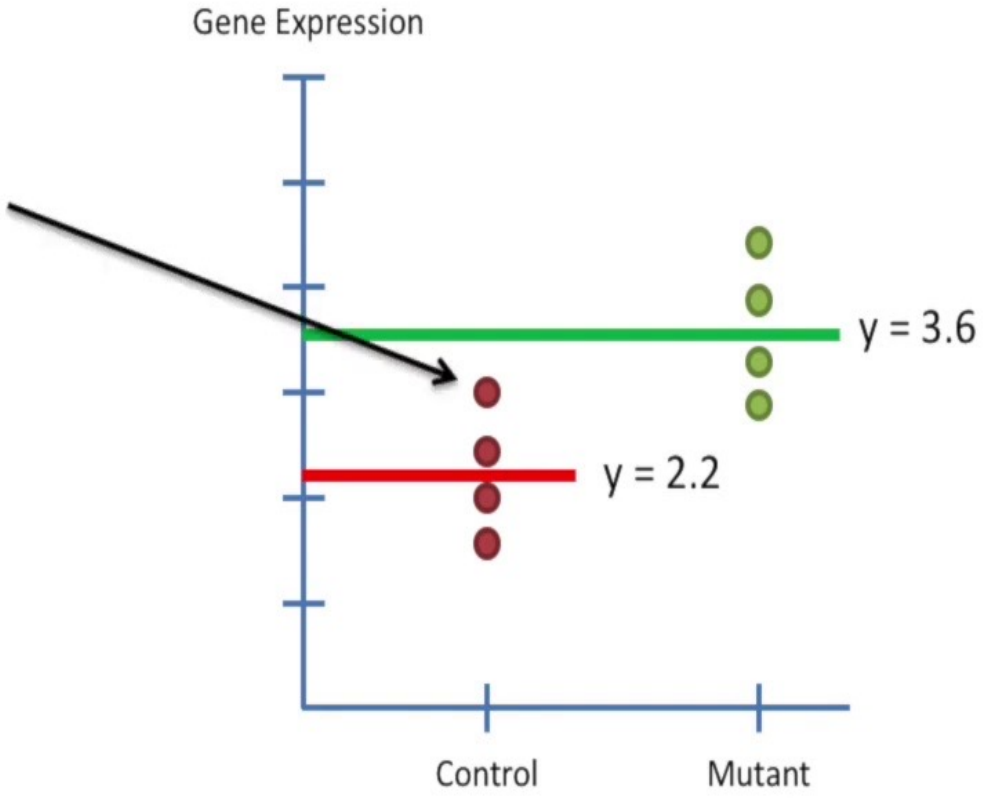
This is going to look weird, but just bear with me.

Keep in mind that the goal is to have a flexible way for a computer to solve this, and every other "least-squares" based problem, without having to create a whole new method each time.



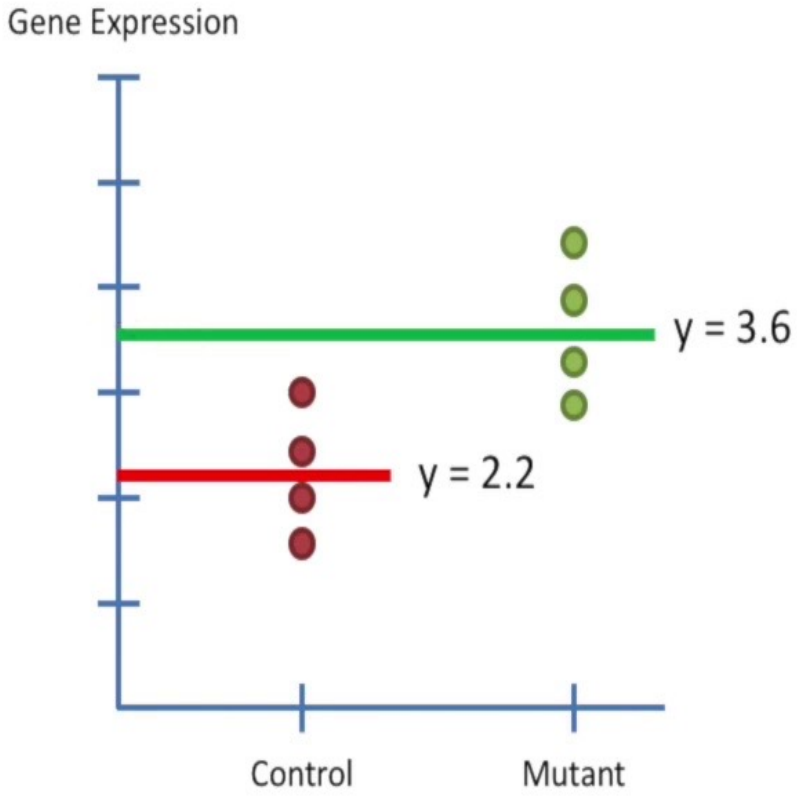
$y = 1 \times 2.2 + 0 \times 3.6 + \text{the residual}$

This is the equation (which combines both lines) for this point



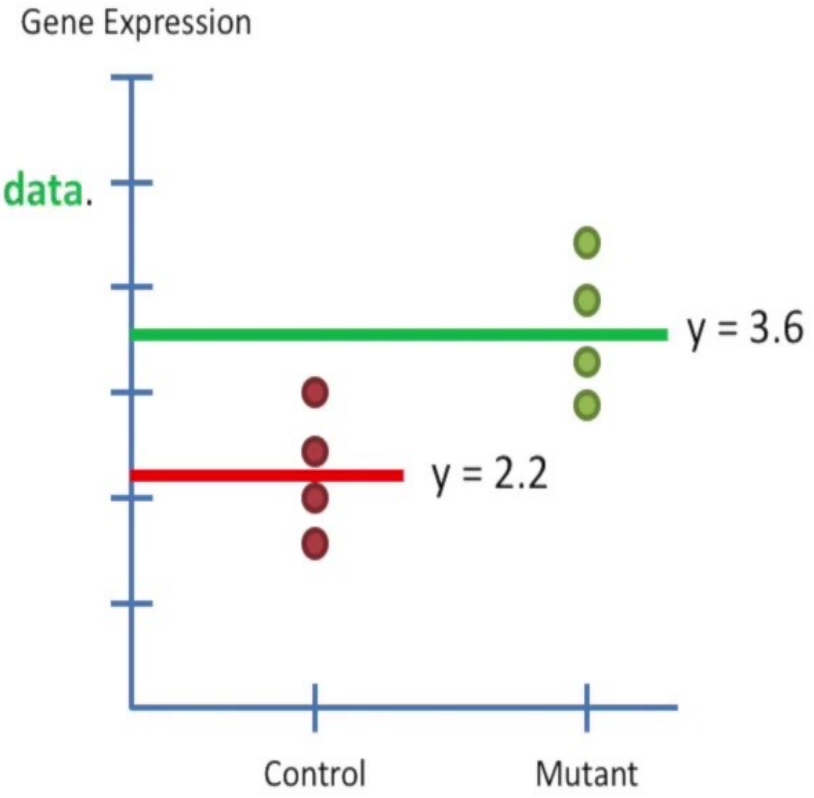
$y = 1 \times 2.2 + 0 \times 3.6 + \text{the residual}$

1 times the mean of the **control data**.

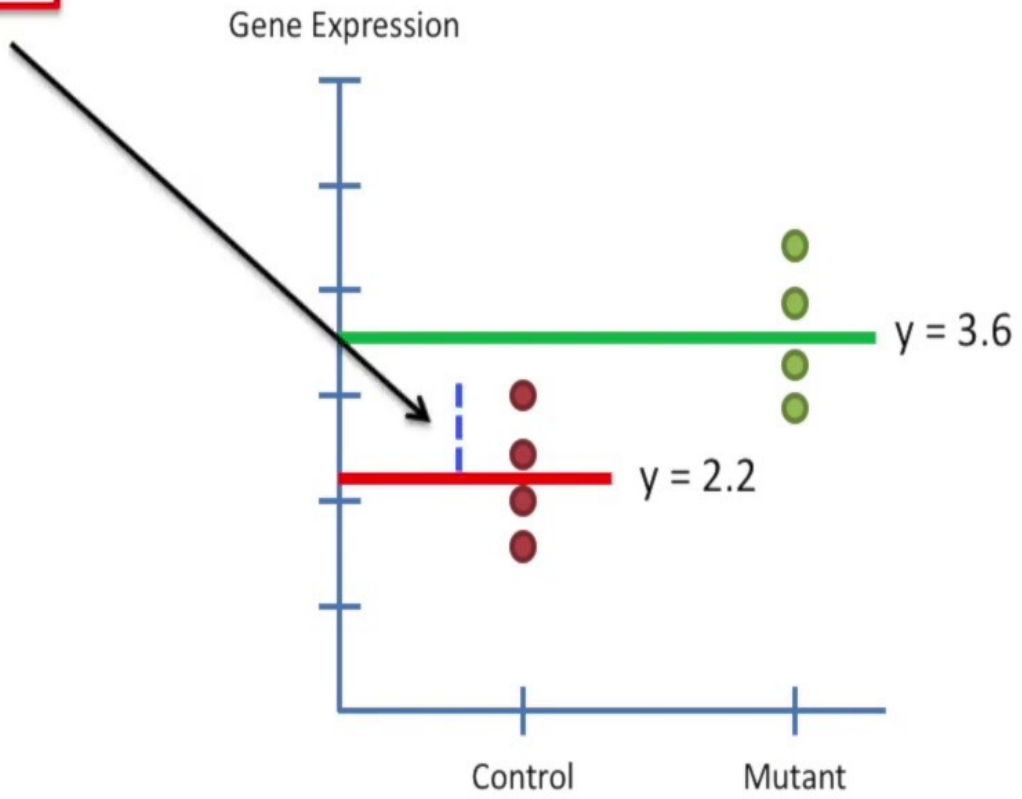


$y = 1 \times 2.2 + 0 \times 3.6 + \text{the residual}$

0 times the mean of the **mutant data**.



$y = 1 \times 2.2 + 0 \times 3.6 +$ the residual



$y = 1 \times 2.2 + 0 \times 3.6 + \text{the residual}$

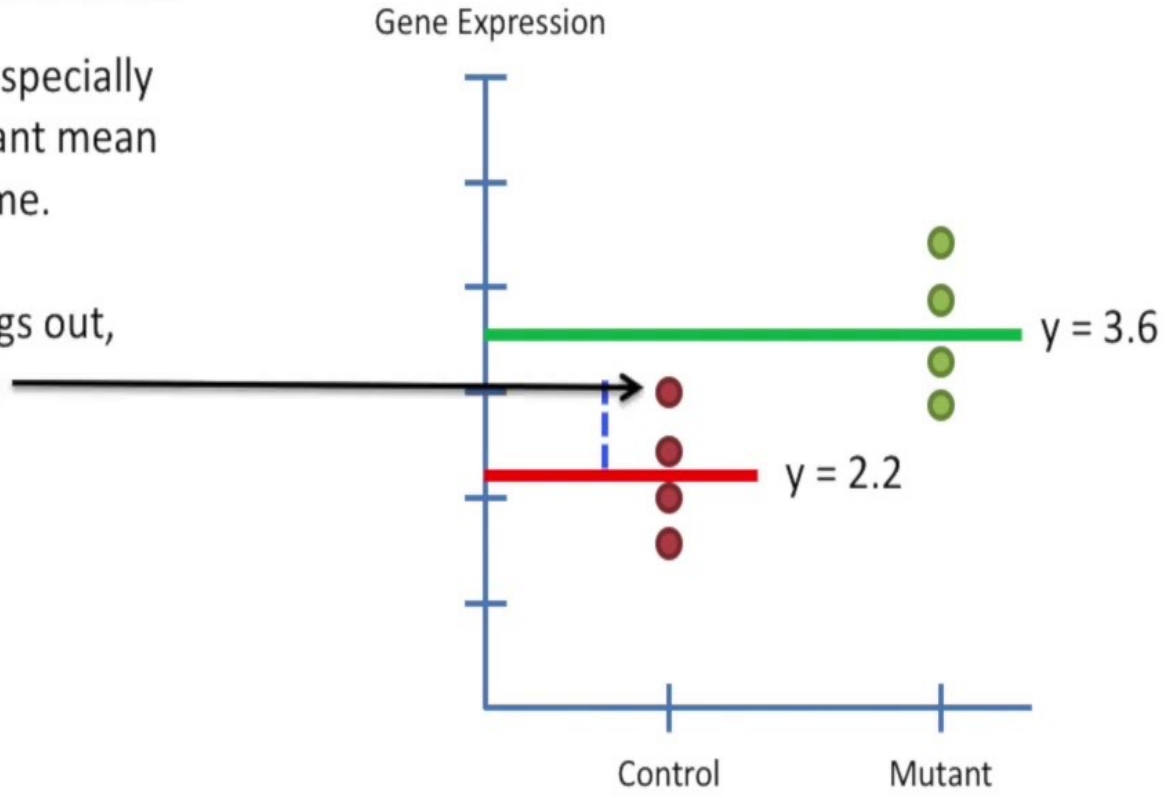
Yes, this is strange, especially multiplying the mutant mean by 0, but bear with me.

$y = 1 \times 2.2 + 0 \times 3.6 + \text{the residual}$

Yes, this is strange, especially multiplying the mutant mean by 0, but bear with me.

If we multiplied things out, the equation for this point would be…

$y = 2.2 + \text{the residual}$



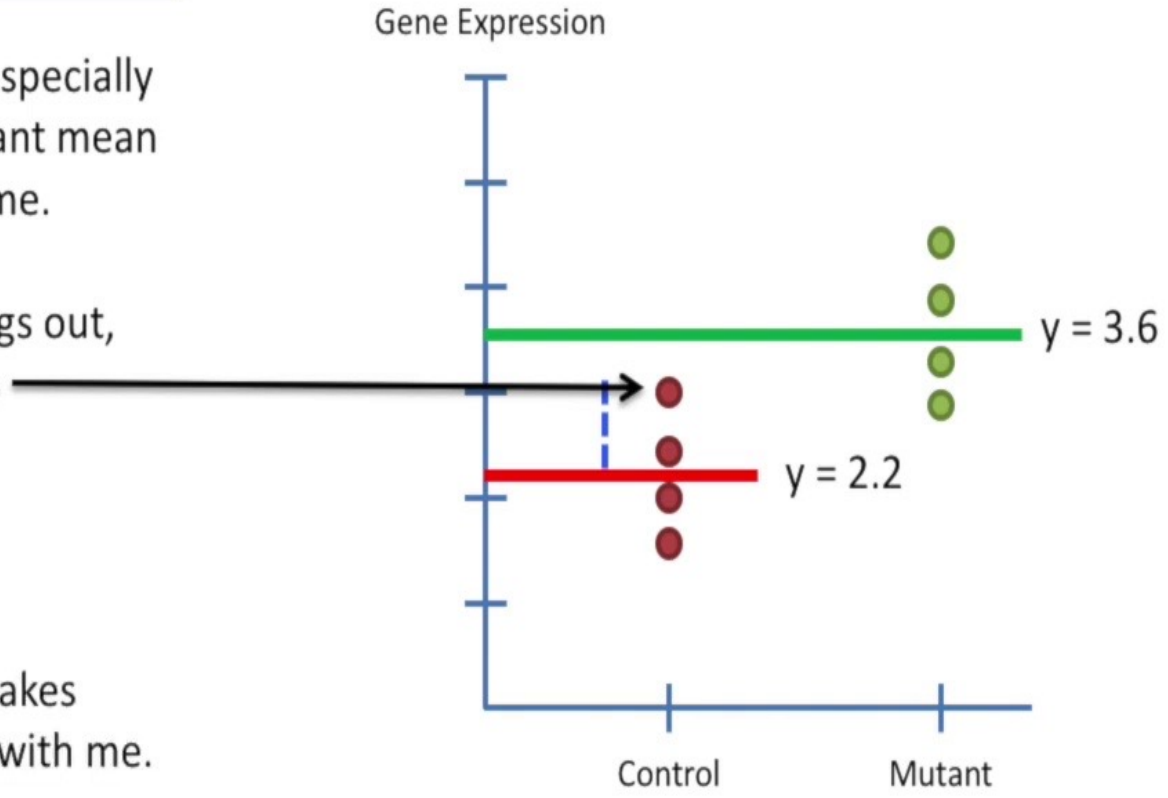
$y = 1 \times 2.2 + 0 \times 3.6 + \text{the residual}$

Yes, this is strange, especially multiplying the mutant mean by 0, but bear with me.

If we multiplied things out, the equation for this point would be…

$y = 2.2 + \text{the residual}$

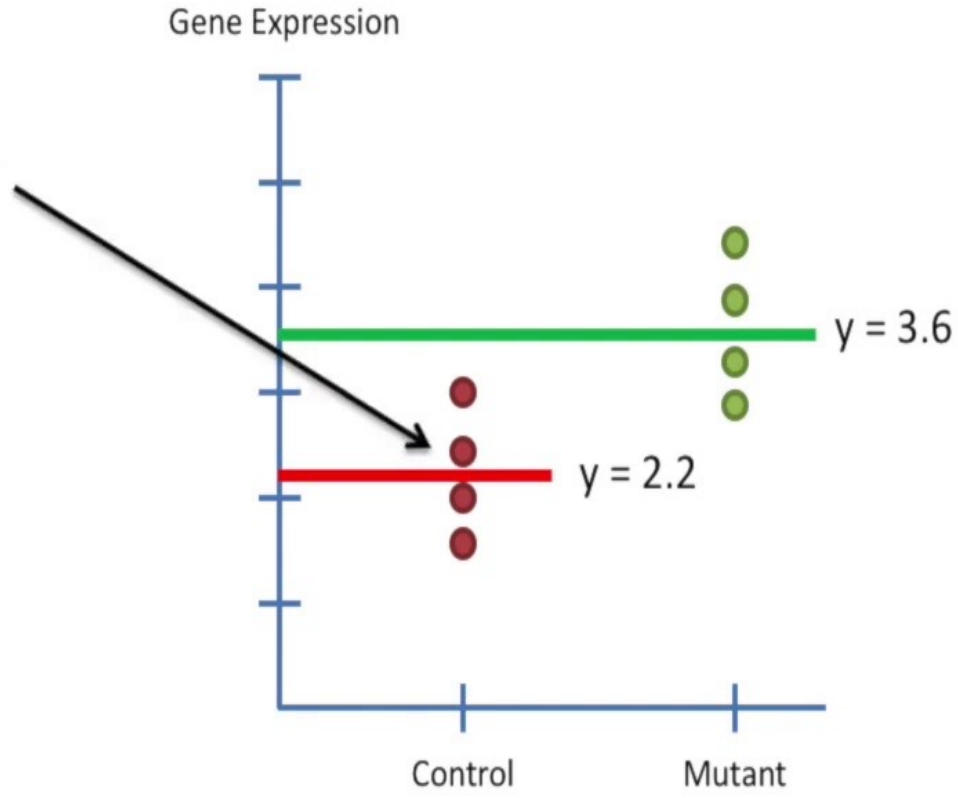
… and that sort of makes sense. But just bear with me.



$y = 1 \times 2.2 + 0 \times 3.6 + \text{the residual}$

$y = 1 \times 2.2 + 0 \times 3.6 + \text{the residual}$

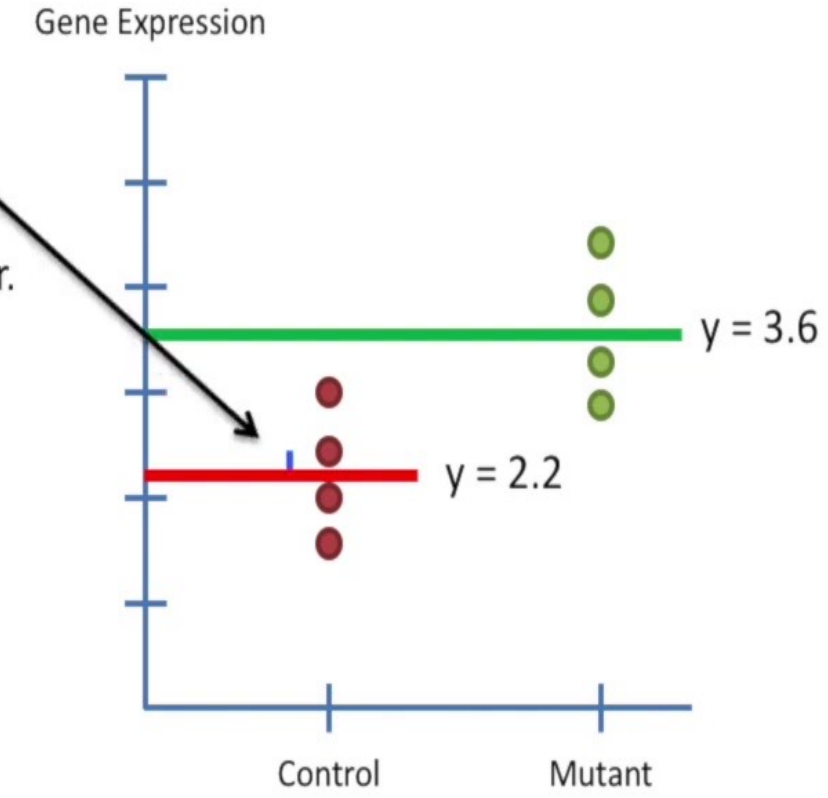
This is the equation for the next point.



$y = 1 \times 2.2 + 0 \times 3.6 + \text{the residual}$

$y = 1 \times 2.2 + 0 \times 3.6 + \text{the residual}$

The only difference is the residual. This one is smaller.

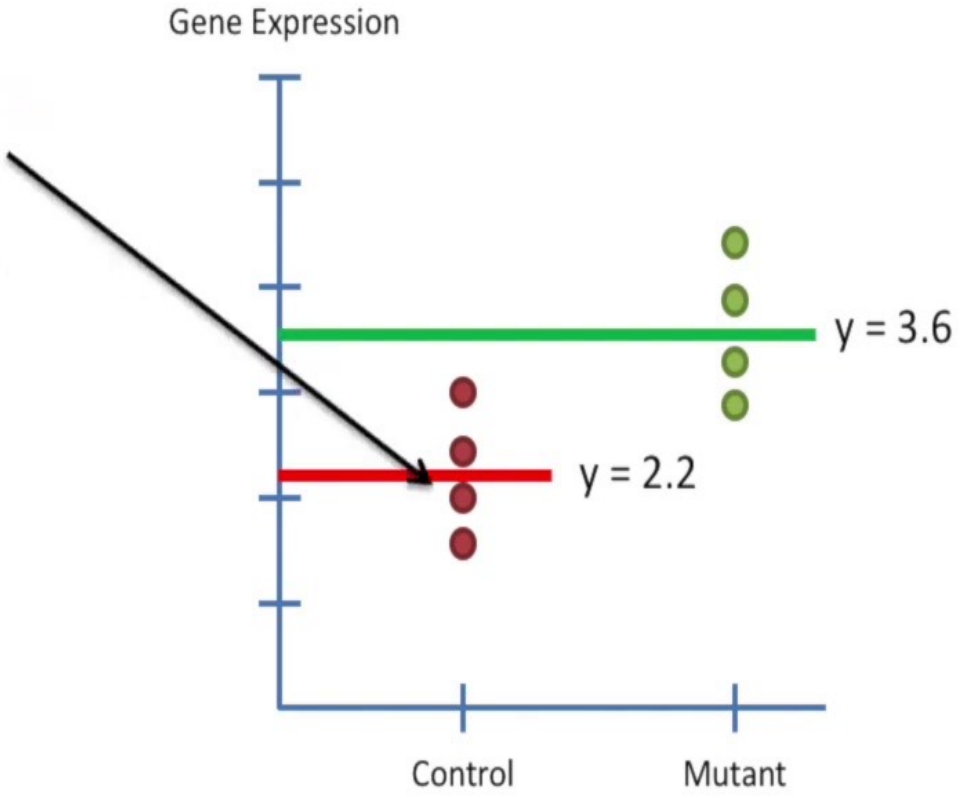


$y = 1 \times 2.2 + 0 \times 3.6 + \text{the residual}$

$y = 1 \times 2.2 + 0 \times 3.6 + \text{the residual}$

$y = 1 \times 2.2 + 0 \times 3.6 + \text{the residual}$

This is the equation for the next point.

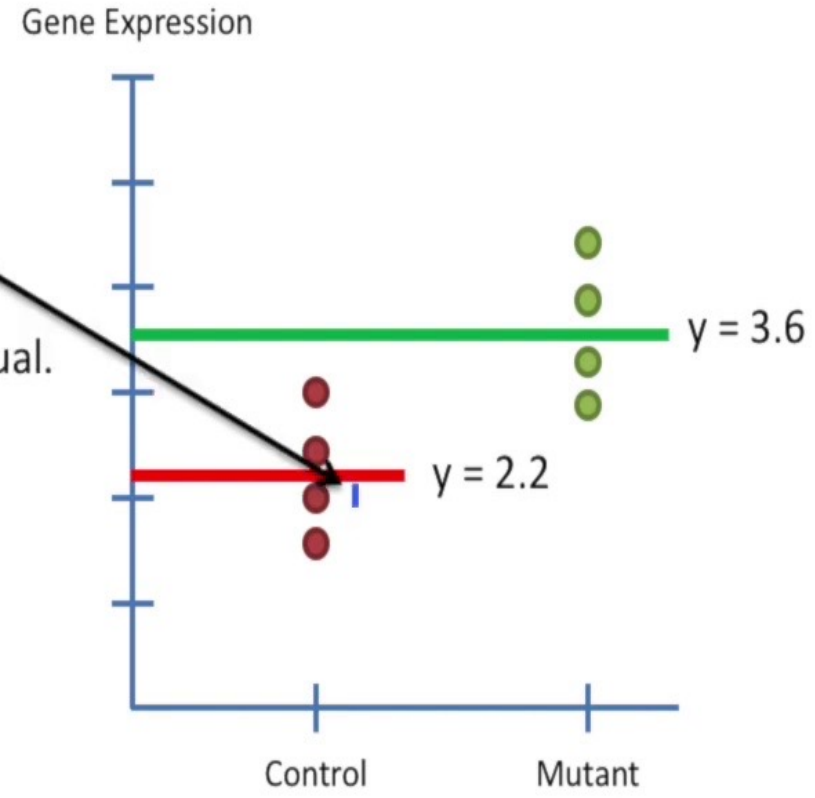


$y = 1 \times 2.2 + 0 \times 3.6 + \text{the residual}$

$y = 1 \times 2.2 + 0 \times 3.6 + \text{the residual}$

$y = 1 \times 2.2 + 0 \times 3.6 + \text{the residual}$

Again, the only difference is the residual.



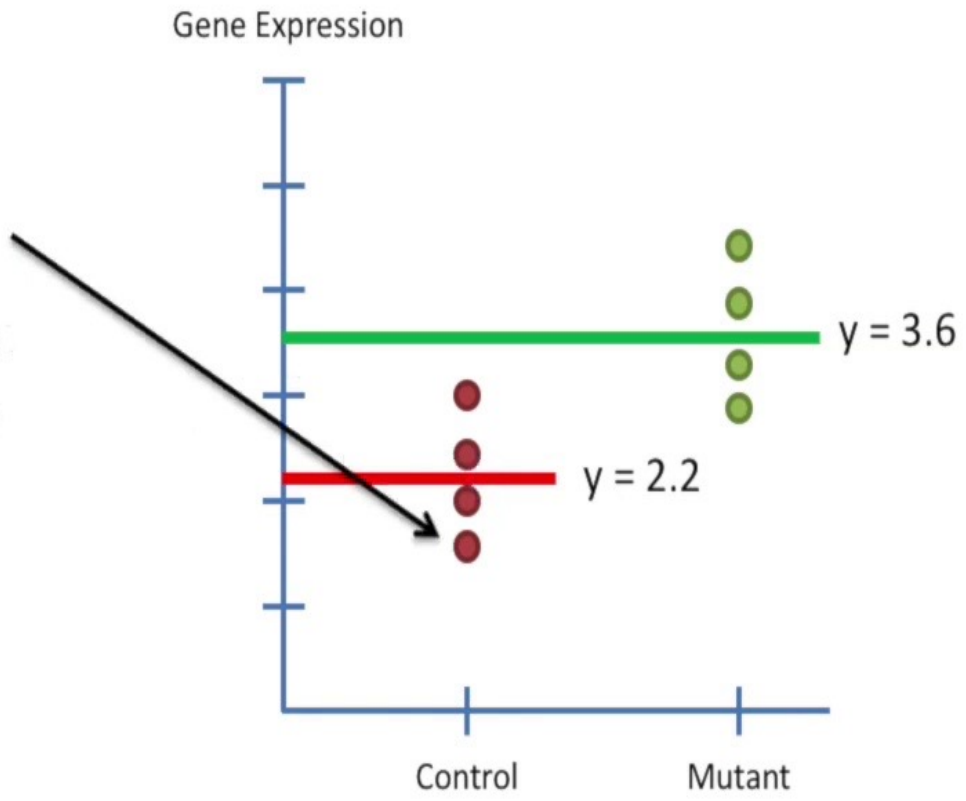
$y = 1 \times 2.2 + 0 \times 3.6 + \text{the residual}$

$y = 1 \times 2.2 + 0 \times 3.6 + \text{the residual}$

$y = 1 \times 2.2 + 0 \times 3.6 + \text{the residual}$

$y = 1 \times 2.2 + 0 \times 3.6 + \text{the residual}$

This is the equation for the next point.



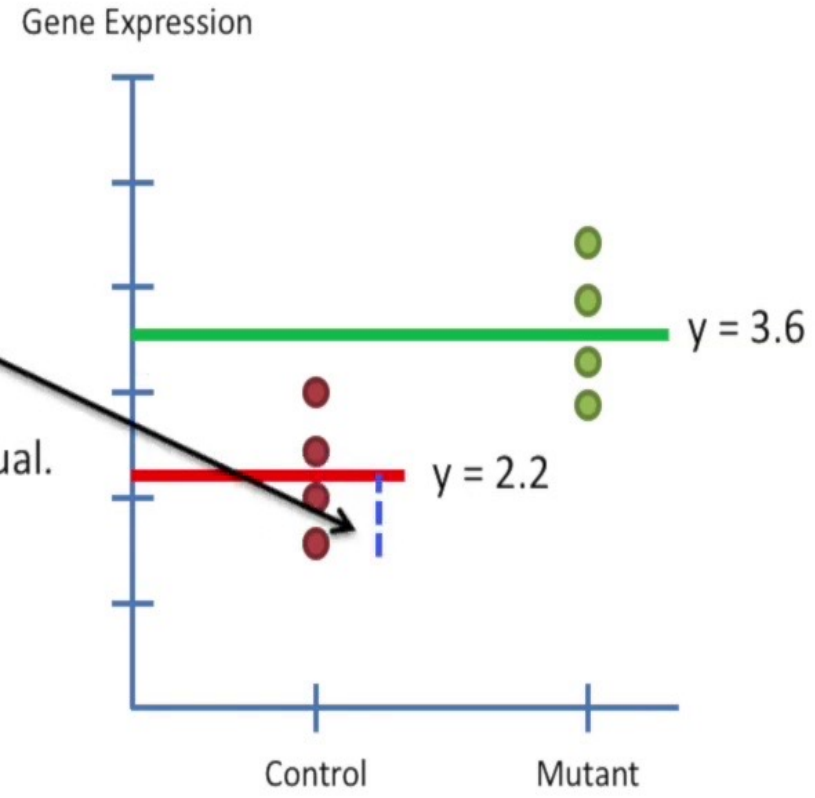
$y = 1 \times 2.2 + 0 \times 3.6 + \text{the residual}$

$y = 1 \times 2.2 + 0 \times 3.6 + \text{the residual}$

$y = 1 \times 2.2 + 0 \times 3.6 + \text{the residual}$

$y = 1 \times 2.2 + 0 \times 3.6 + \text{the residual}$

Again, the only difference is the residual.



$y = 1 \times 2.2 + 0 \times 3.6 + \text{the residual}$

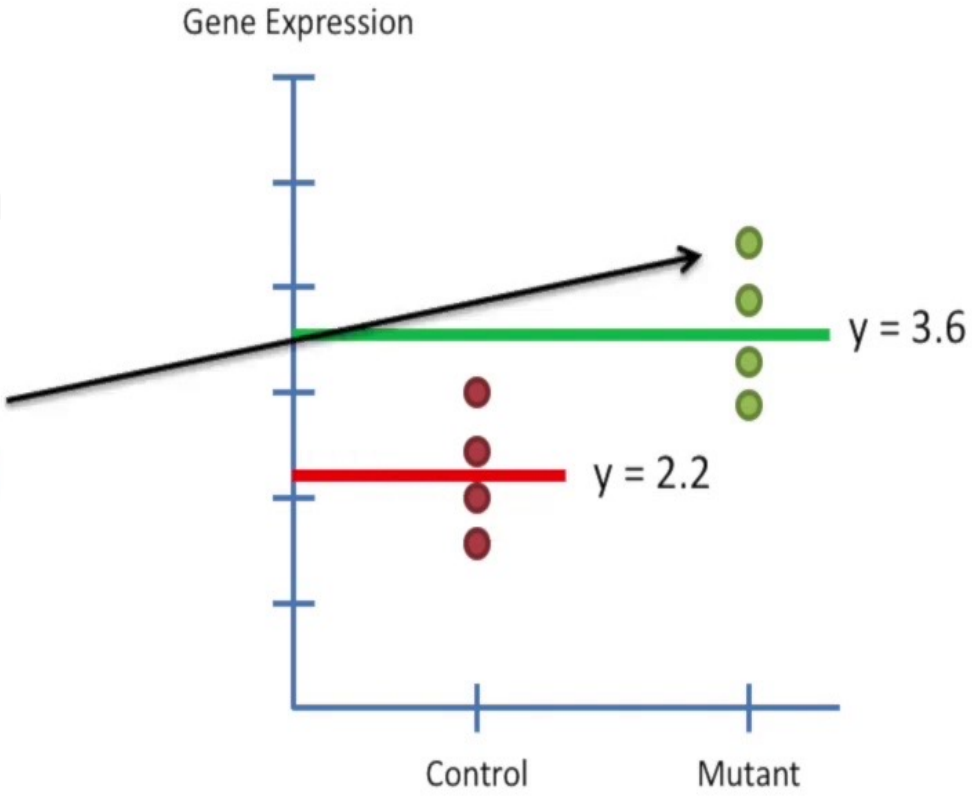
$y = 1 \times 2.2 + 0 \times 3.6 + \text{the residual}$

$y = 1 \times 2.2 + 0 \times 3.6 + \text{the residual}$

$y = 1 \times 2.2 + 0 \times 3.6 + \text{the residual}$

$y = 0 \times 2.2 + 1 \times 3.6 + \text{the residual}$

This is the equation for the first point in the mutant dataset.



$y = 1 \times 2.2 + 0 \times 3.6 + \text{the residual}$

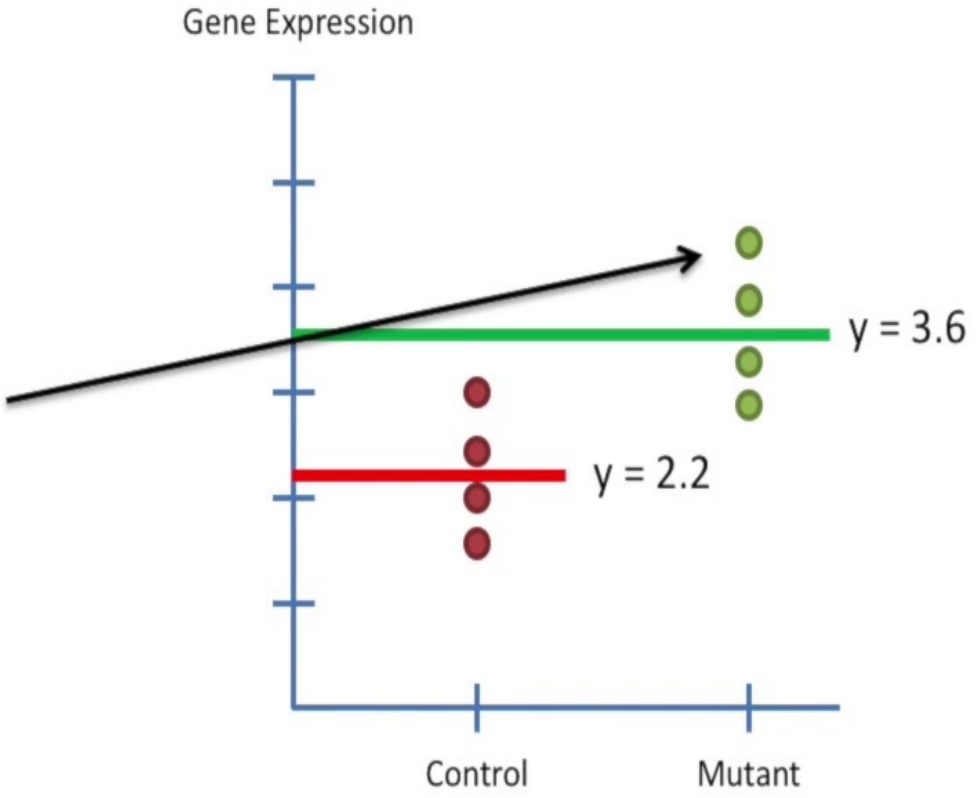
$y = 1 \times 2.2 + 0 \times 3.6 + \text{the residual}$

$y = 1 \times 2.2 + 0 \times 3.6 + \text{the residual}$

$y = 1 \times 2.2 + 0 \times 3.6 + \text{the residual}$

$y = 0 \times 2.2 + 1 \times 3.6 + \text{the residual}$

Now we are multiplying the **control** mean by 0…



$y = 1 \times 2.2 + 0 \times 3.6 + \text{the residual}$

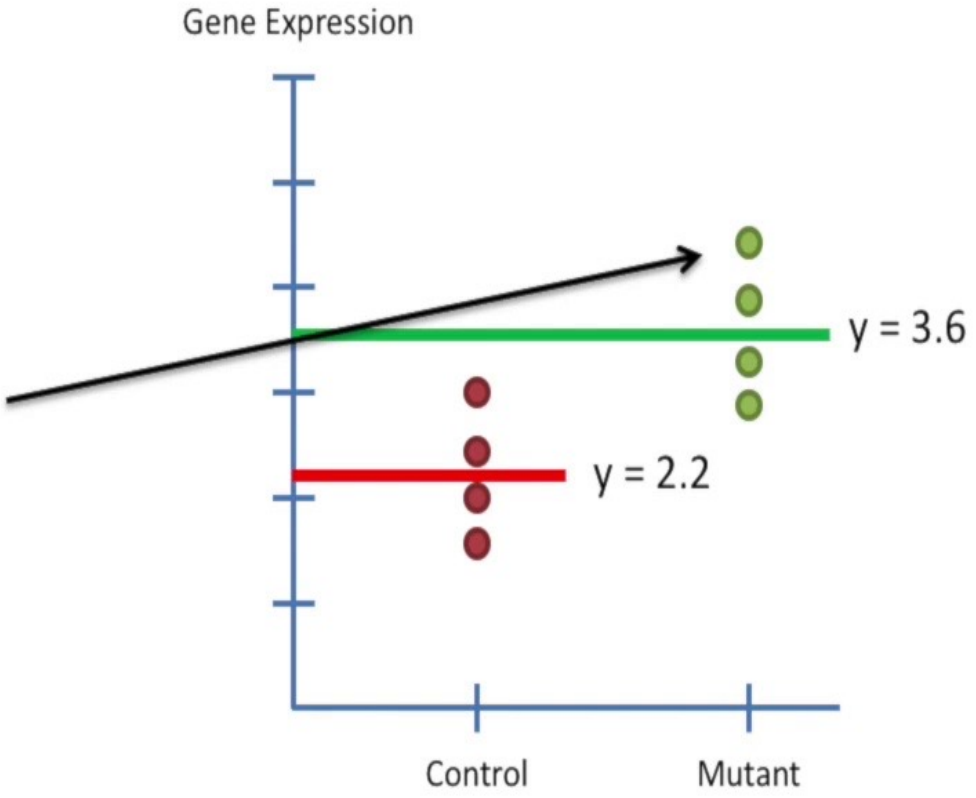
$y = 1 \times 2.2 + 0 \times 3.6 + \text{the residual}$

$y = 1 \times 2.2 + 0 \times 3.6 + \text{the residual}$

$y = 1 \times 2.2 + 0 \times 3.6 + \text{the residual}$

$y = 0 \times 2.2 + 1 \times 3.6 + \text{the residual}$

...and multiplying the **mutant** mean by 1.



$y = 1 \times 2.2 + 0 \times 3.6 + \text{the residual}$

$y = 1 \times 2.2 + 0 \times 3.6 + \text{the residual}$

$y = 1 \times 2.2 + 0 \times 3.6 + \text{the residual}$

$y = 1 \times 2.2 + 0 \times 3.6 + \text{the residual}$

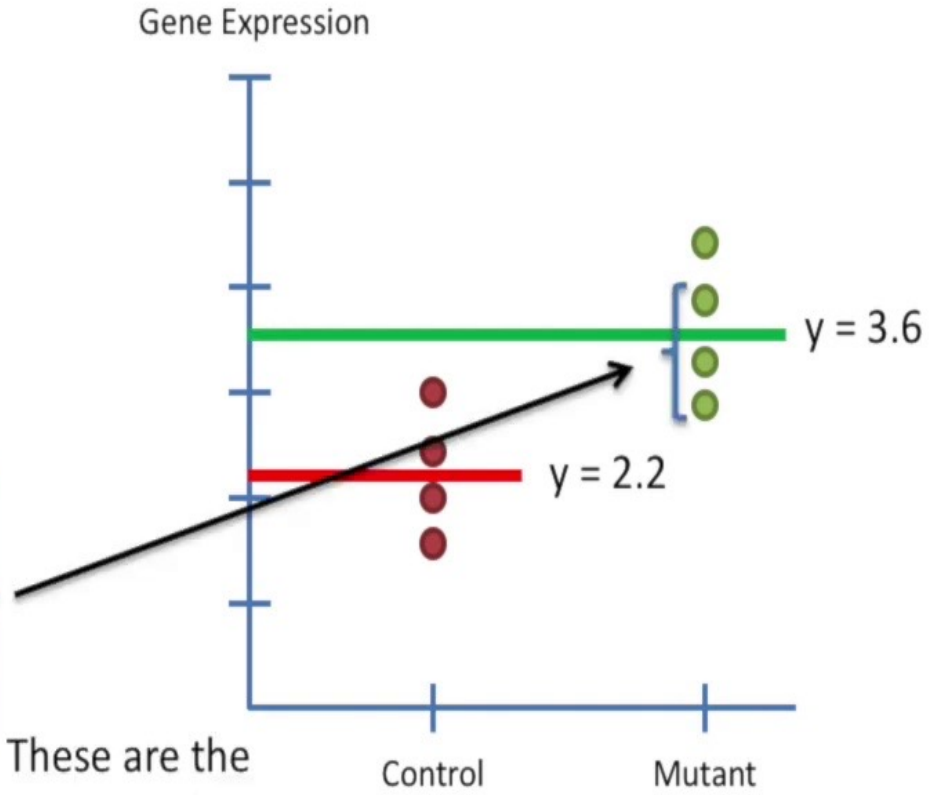
$y = 0 \times 2.2 + 1 \times 3.6 + \text{the residual}$

$y = 0 \times 2.2 + 1 \times 3.6 + \text{the residual}$

$y = 0 \times 2.2 + 1 \times 3.6 + \text{the residual}$

$y = 0 \times 2.2 + 1 \times 3.6 + \text{the residual}$

These are the equations for the remaining points.



$y = 1 \times 2.2 + 0 \times 3.6 + \text{the residual}$

$y = 1 \times 2.2 + 0 \times 3.6 + \text{the residual}$

$y = 1 \times 2.2 + 0 \times 3.6 + \text{the residual}$

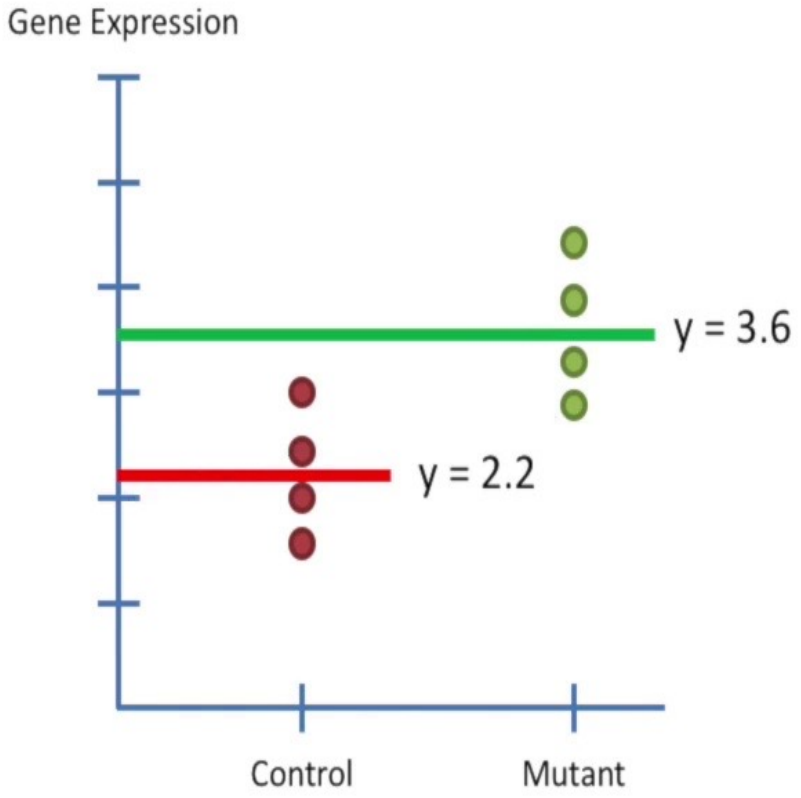
$y = 1 \times 2.2 + 0 \times 3.6 + \text{the residual}$

$y = 0 \times 2.2 + 1 \times 3.6 + \text{the residual}$

$y = 0 \times 2.2 + 1 \times 3.6 + \text{the residual}$

$y = 0 \times 2.2 + 1 \times 3.6 + \text{the residual}$

$y = 0 \times 2.2 + 1 \times 3.6 + \text{the residual}$



Now let's focus on the 1's and 0's.

$y = 1 \times 2.2 + 0 \times 3.6 + \text{the residual}$

$y = 1 \times 2.2 + 0 \times 3.6 + \text{the residual}$

$y = 1 \times 2.2 + 0 \times 3.6 + \text{the residual}$

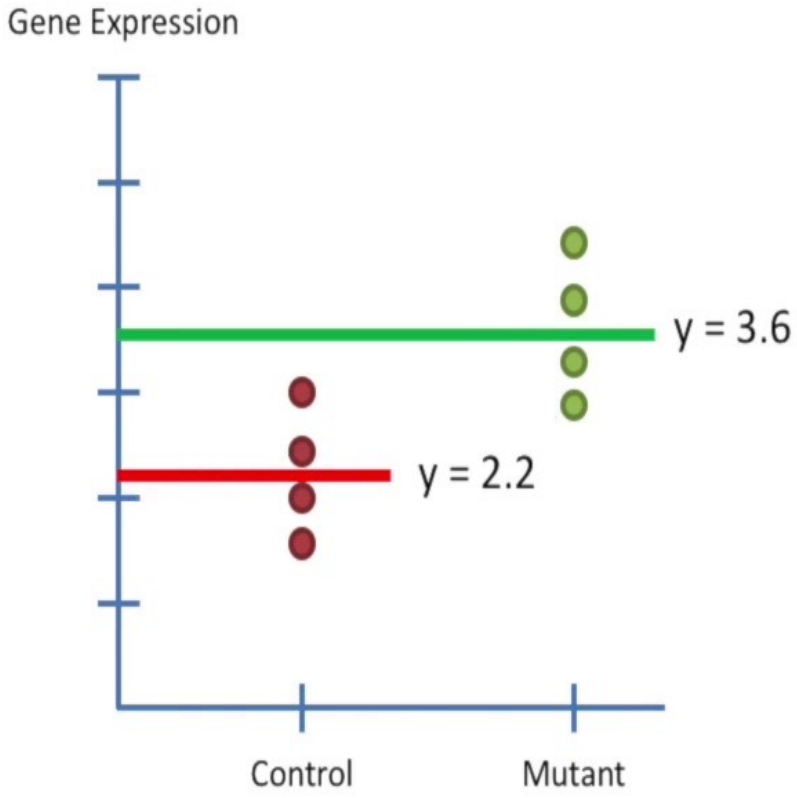
$y = 1 \times 2.2 + 0 \times 3.6 + \text{the residual}$

$y = 0 \times 2.2 + 1 \times 3.6 + \text{the residual}$

$y = 0 \times 2.2 + 1 \times 3.6 + \text{the residual}$

$y = 0 \times 2.2 + 1 \times 3.6 + \text{the residual}$

$y = 0 \times 2.2 + 1 \times 3.6 + \text{the residual}$



They function like on/off switches for the two means.

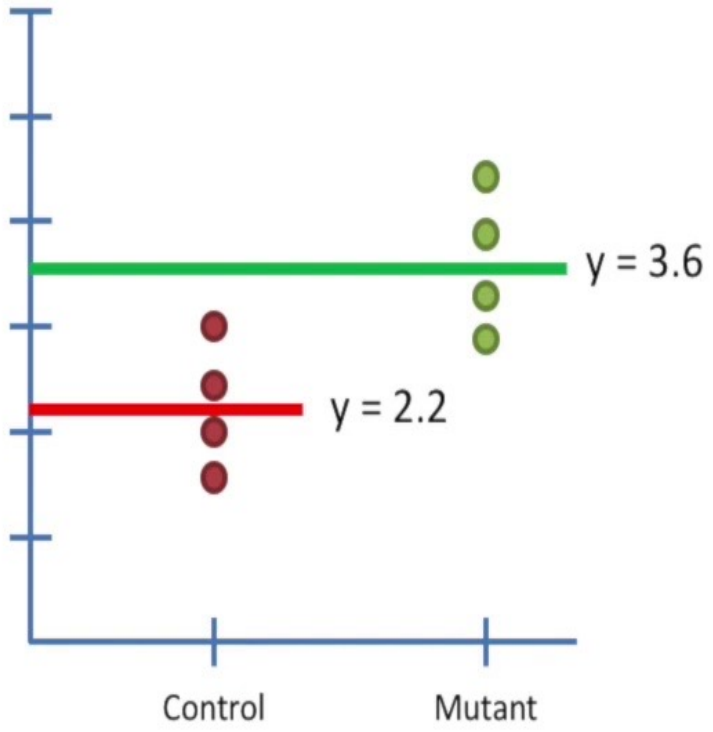
1	0
1	0
1	0
1	0
0	1
0	1
0	1
0	1

When we isolate the 1s and 0s, they form a matrix called a “design matrix”.

1	0
1	0
1	0
1	0
0	1
0	1
0	1
0	1

The design matrix can be combined with an abstract version of the equation to represent a "fit" to the data.

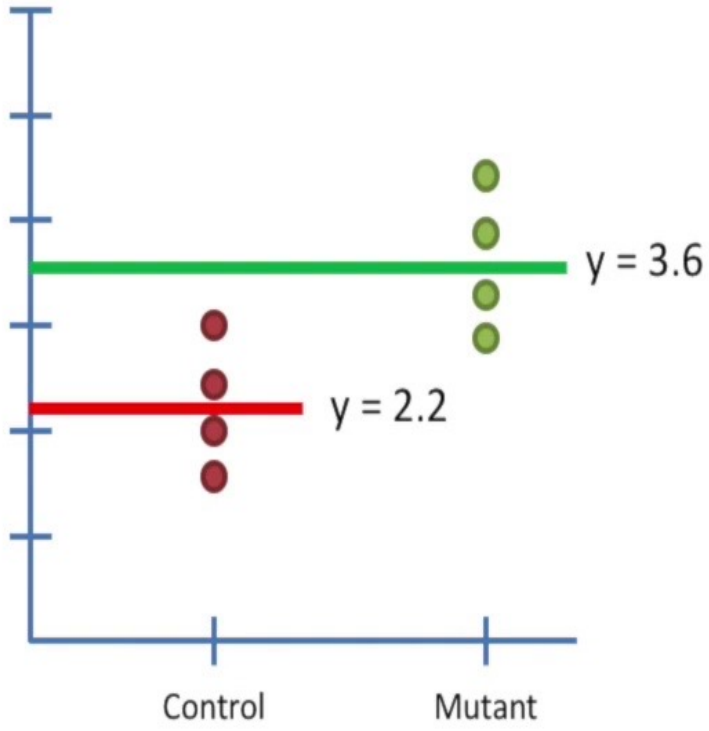
$y = column1 \times \mathbf{2.2} + column2 \times \mathbf{3.6}$



1	0
1	0
1	0
1	0
0	1
0	1
0	1
0	1

$y = \text{column1} \times \mathbf{2.2} + \text{column2} \times \mathbf{3.6}$

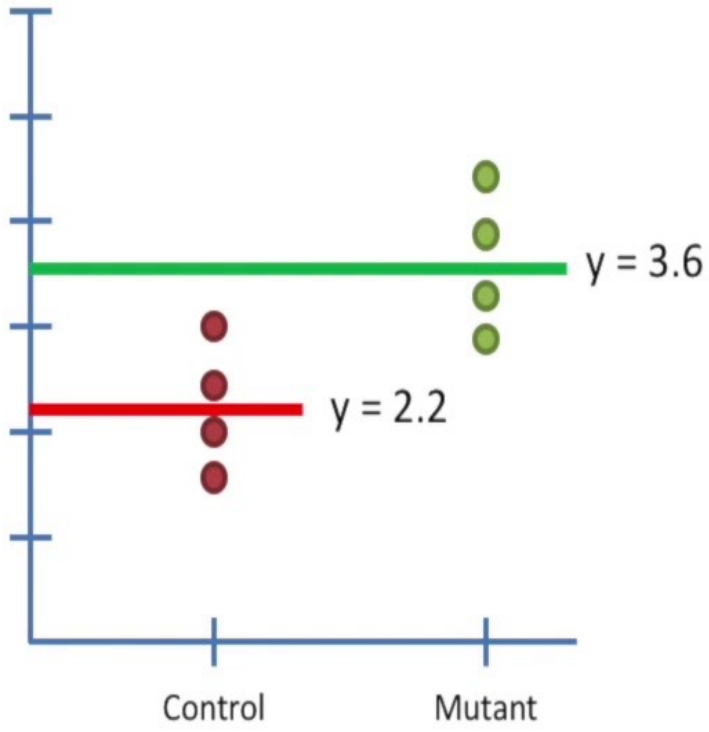
Column1 turns the **control** mean “on” or “off”.



1	0
1	0
1	0
1	0
0	1
0	1
0	1
0	1

$y = \text{column1} \times \mathbf{2.2} + \text{column2} \times \mathbf{3.6}$

Column2 turns the **mutant** mean "on" or "off".

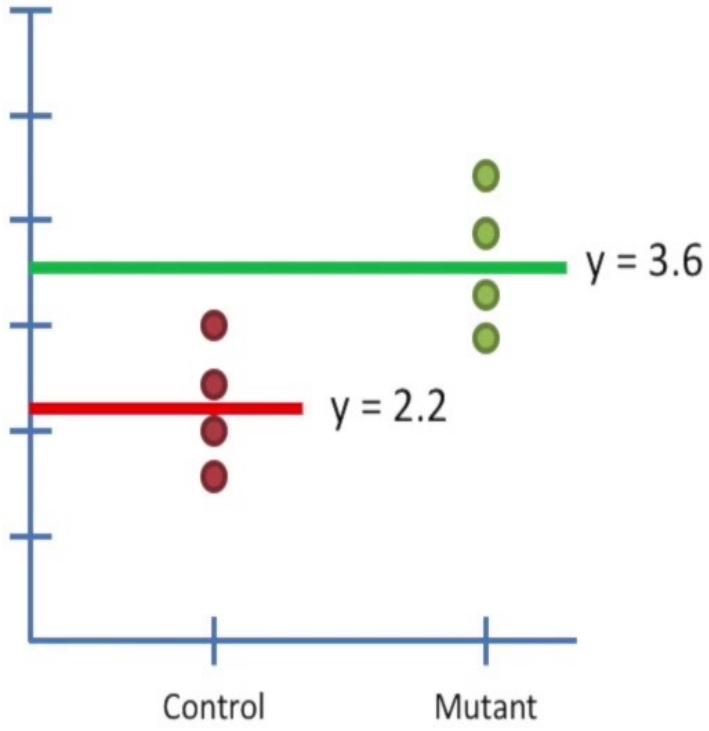


1	0
1	0
1	0
1	0
0	1
0	1
0	1
0	1

$y = \text{column1} \times \mathbf{2.2} + \text{column2} \times \mathbf{3.6}$

In practice, the role of each column is assumed, and the equation is written out like this:

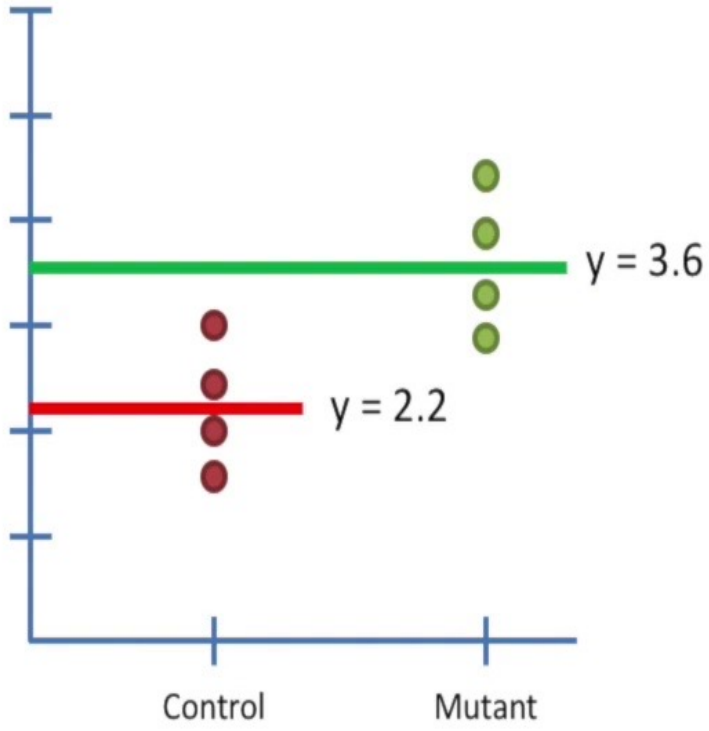
$y = \mathbf{mean}_{\mathbf{control}} + \mathbf{mean}_{\mathbf{mutant}}$



1	0
1	0
1	0
1	0
0	1
0	1
0	1
0	1

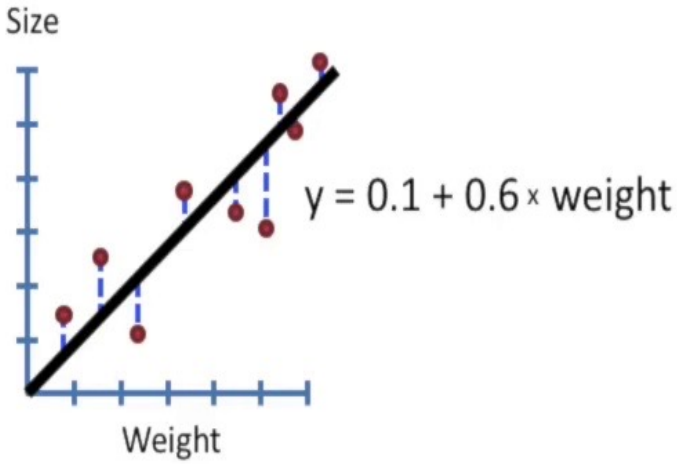
Now that we have the "fit" for the control and mutant data down to a single equation (plus design matrix). We can move on to calculating *F* and the p-value.

$y = \mathbf{mean}_{control} + \mathbf{mean}_{mutant}$



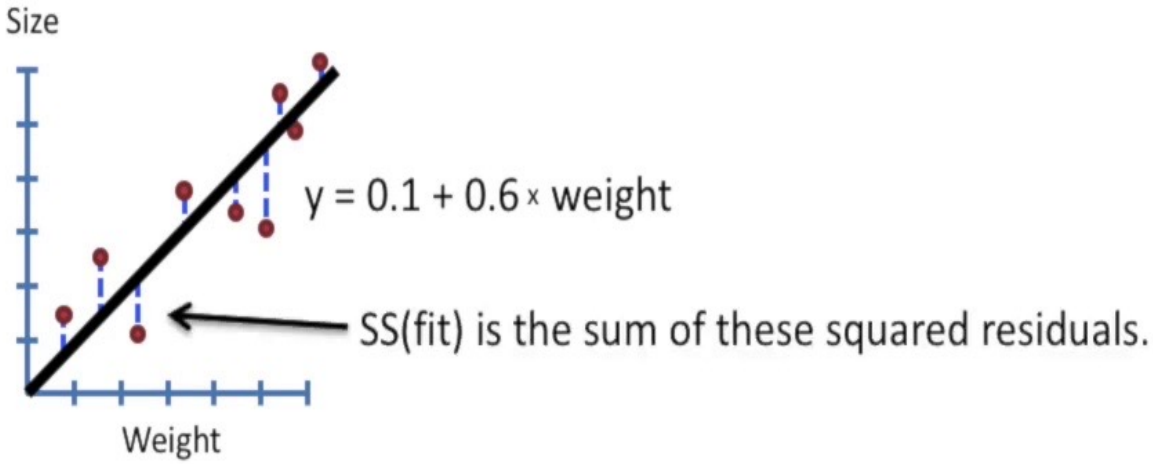
Step 4: Calculate SS(fit), the sum of squares of the residuals around the fitted line(s)

Linear Regression



Step 4: Calculate SS(fit), the sum of squares of the residuals around the fitted line(s)

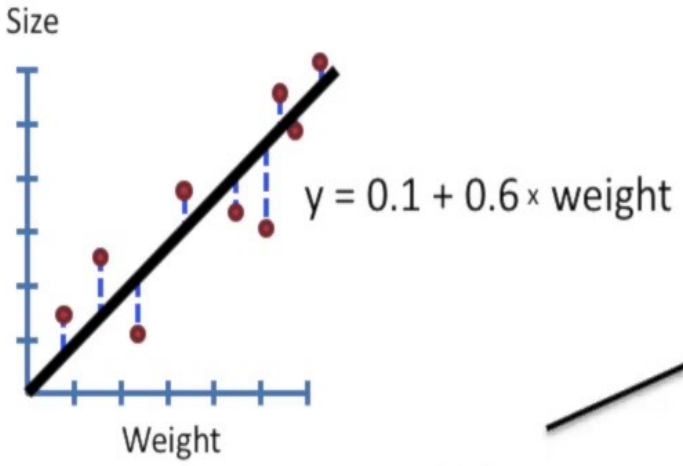
Linear Regression



Step 4: Calculate SS(fit), the sum of squares of the residuals around the fitted line(s)

t-test

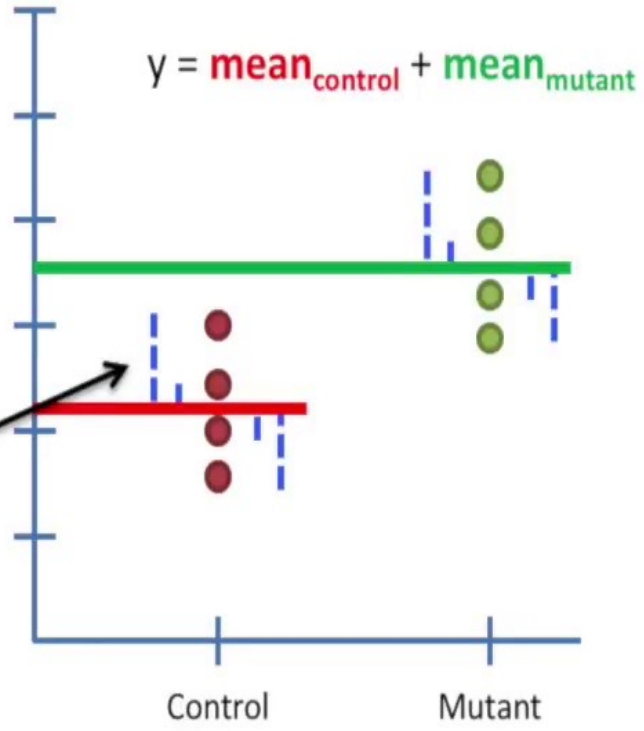
Linear Regression



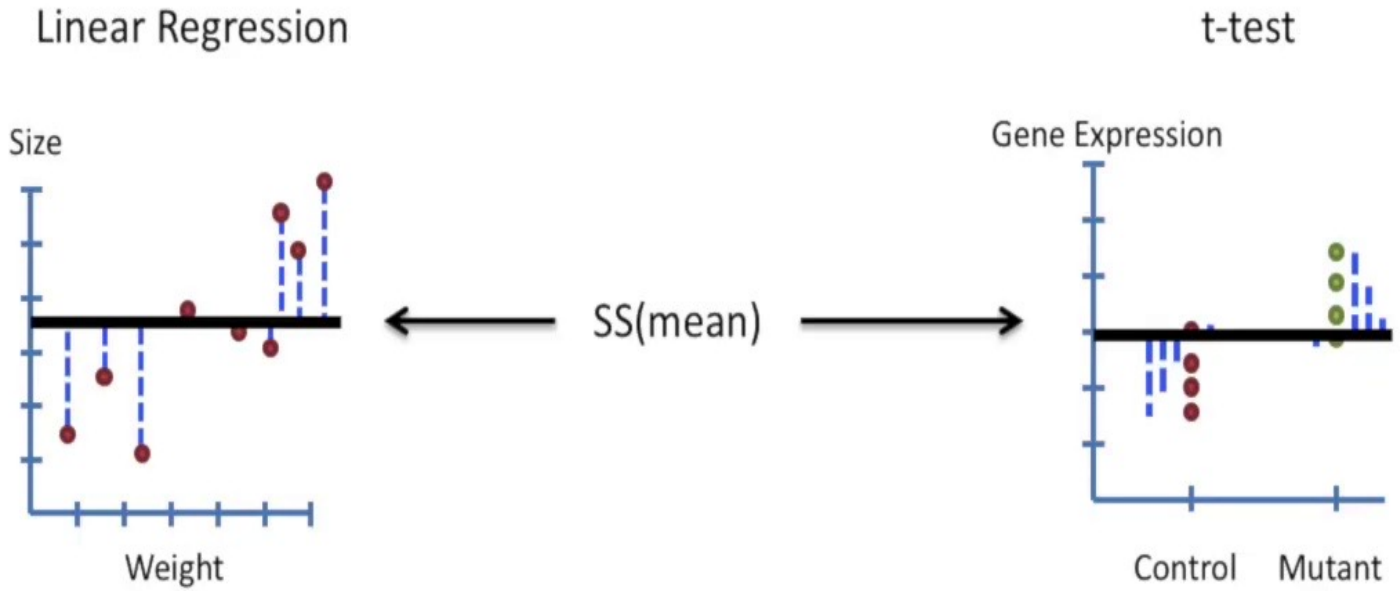
$y = 0.1 + 0.6 \times \text{weight}$

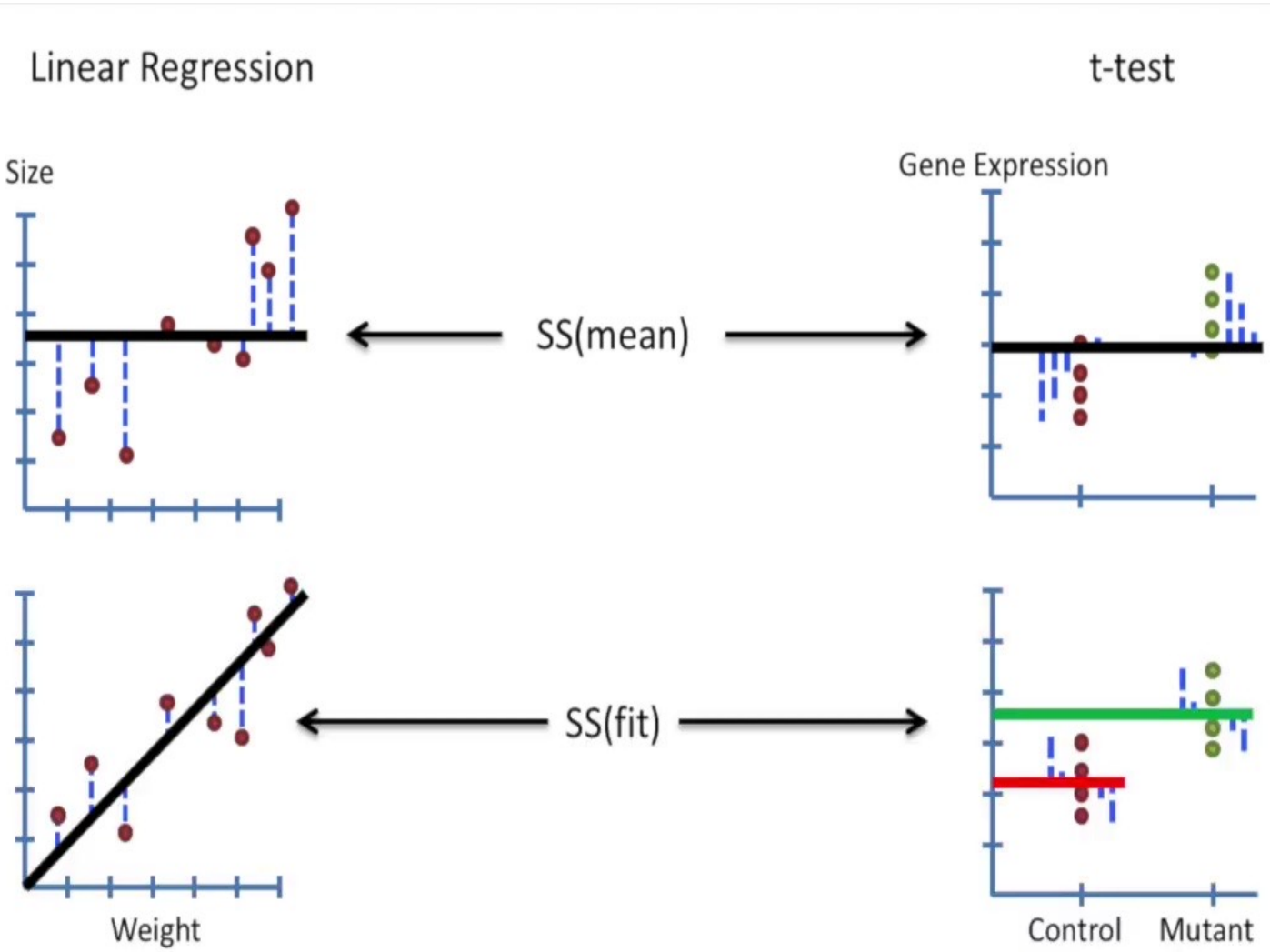
Gene Expression

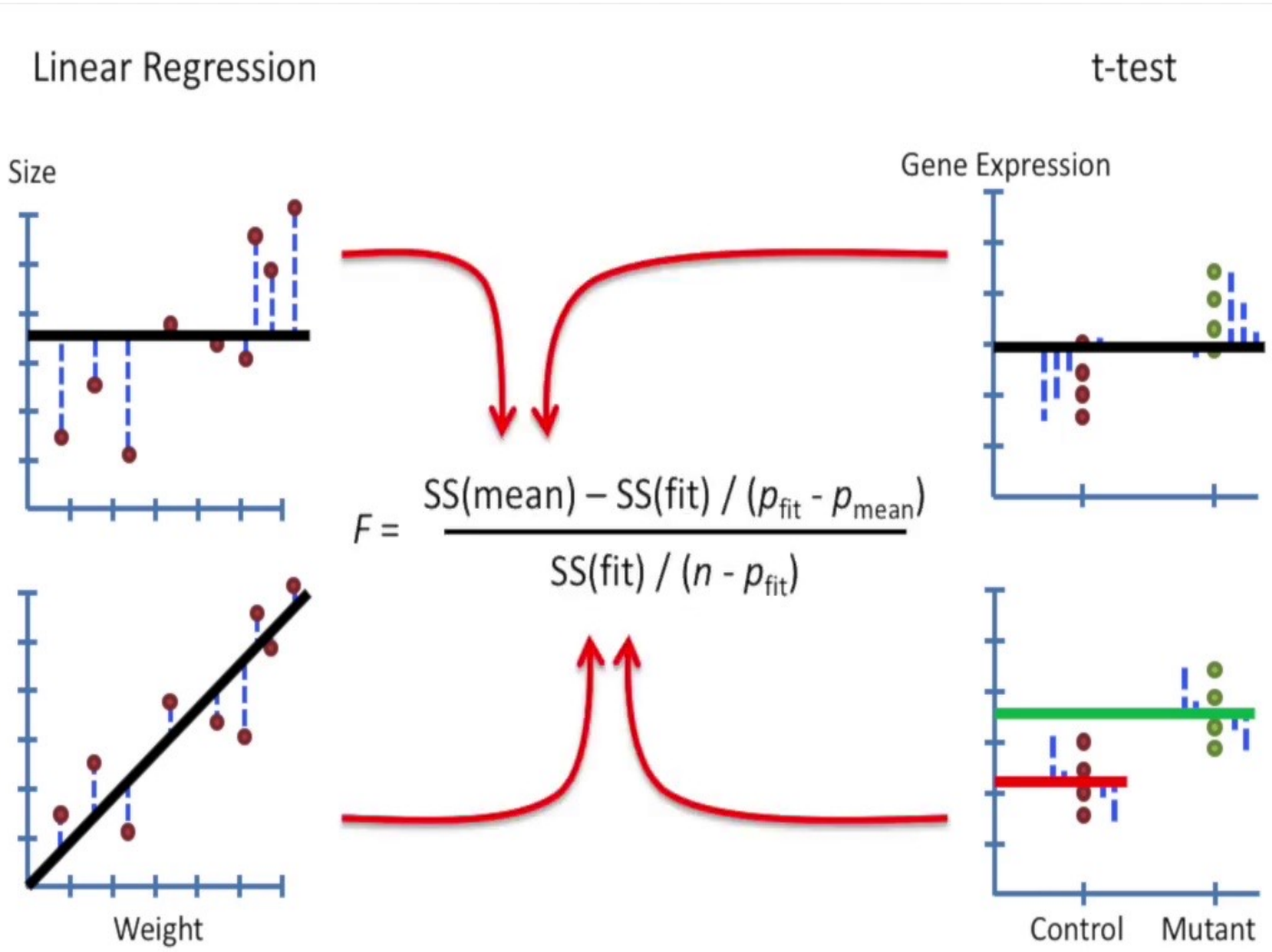
$y = \text{mean}_{\text{control}} + \text{mean}_{\text{mutant}}$



SS(fit) for the t-test is the sum of these squared residuals.

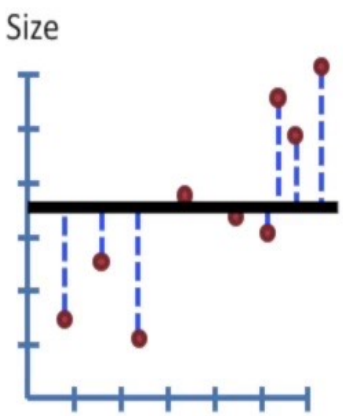

Linear Regression
t-test
Size
Gene Expression
SS(mean)
Weight
Control
Mutant


Linear Regression
t-test
Size
Gene Expression
SS(mean)
SS(fit)
Weight
Control
Mutant


Linear Regression
t-test
Size
Gene Expression
$F = \frac{SS(mean) - SS(fit) / (p_{fit} - p_{mean})}{SS(fit) / (n - p_{fit})}$
Weight
Control
Mutant

Linear Regression

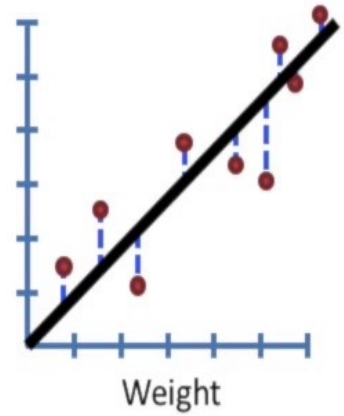
t-test

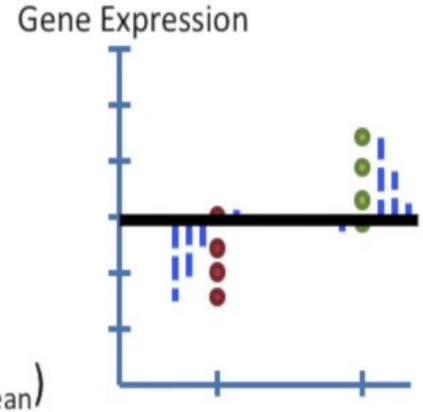


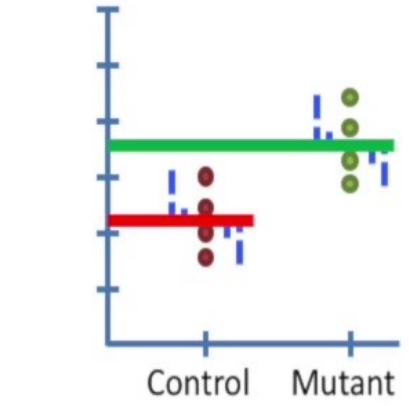
y = mean mouse size

$p_{mean} = 1$

$$F = \frac{SS(mean) - SS(fit) / (p_{fit} - p_{mean})}{SS(fit) / (n - p_{fit})}$$

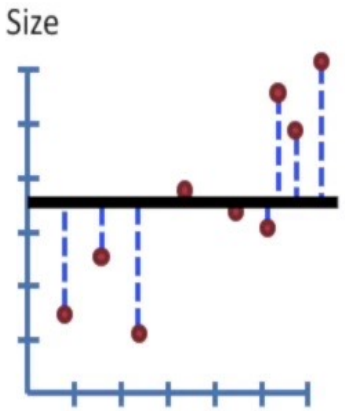


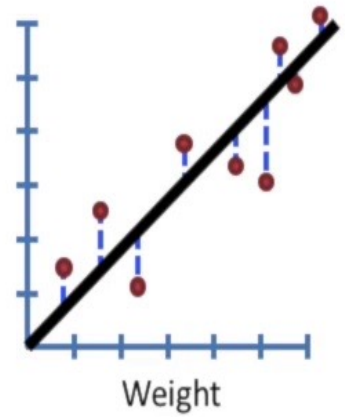




Linear Regression

t-test

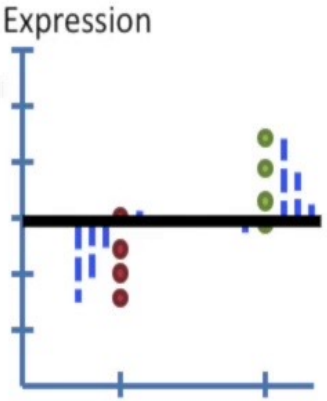


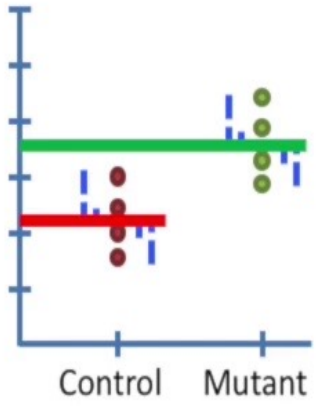


y = mean gene expression

$p_{mean} = 1$

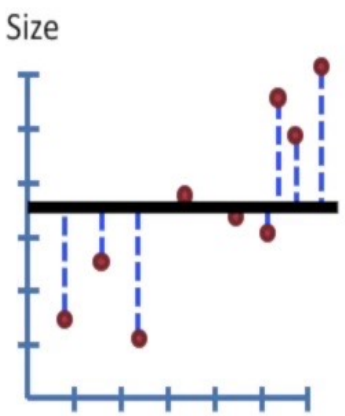
$$F = \frac{SS(mean) - SS(fit) \,/\, (p_{fit} - p_{mean})}{SS(fit) \,/\, (n - p_{fit})}$$

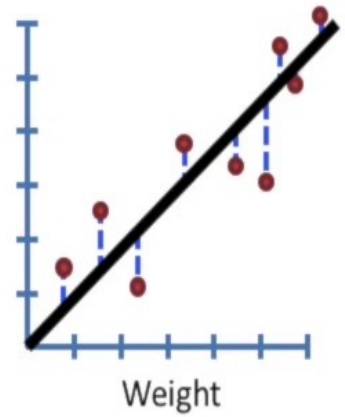




Linear Regression

t-test

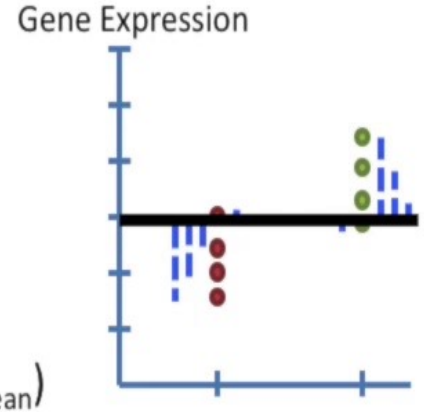


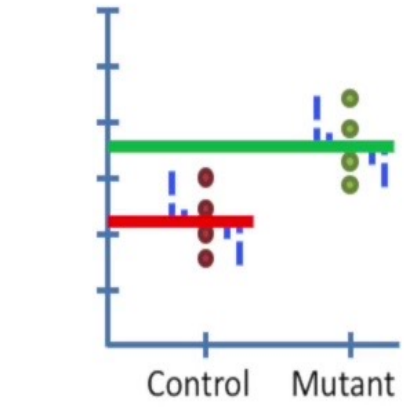


$$F = \frac{SS(mean) - SS(fit) / (p_{fit} - p_{mean})}{SS(fit) / (n - p_{fit})}$$

y = intercept + slope

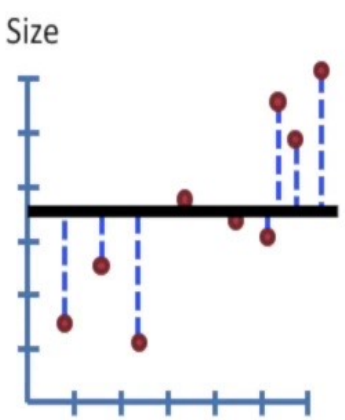
$p_{fit} = 2$

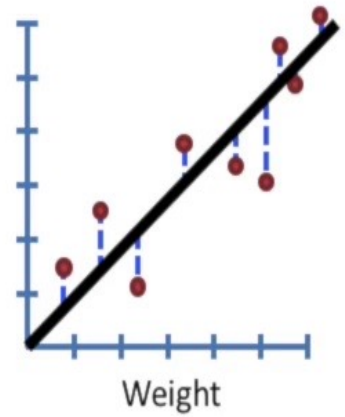




Linear Regression

t-test

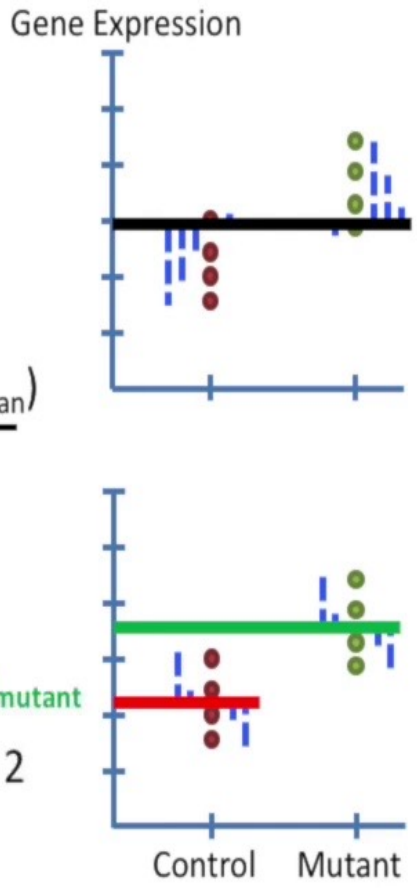




$$F = \frac{SS(mean) - SS(fit) / (p_{fit} - p_{mean})}{SS(fit) / (n - p_{fit})}$$

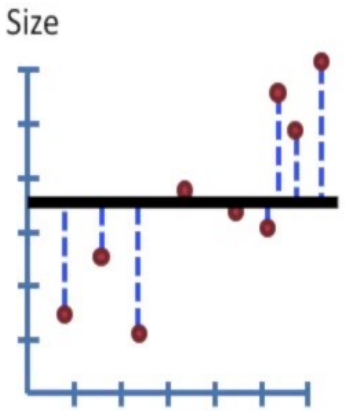
$y = mean_{control} + mean_{mutant}$

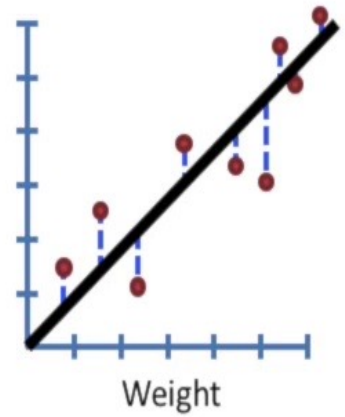
$p_{fit} = 2$



Linear Regression

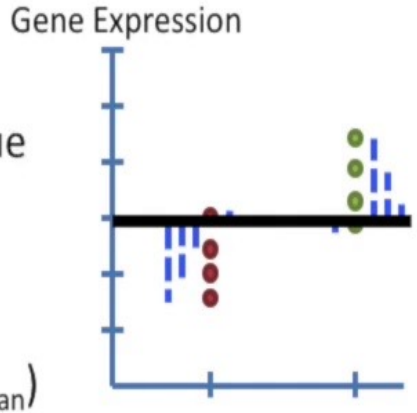
t-test

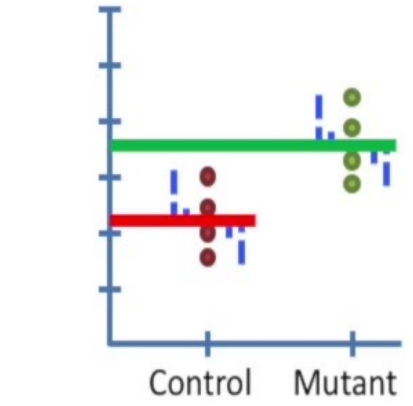




Now we can calculate a p-value for the t-test!

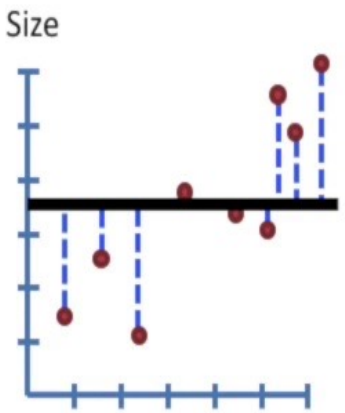
$$F = \frac{SS(mean) - SS(fit) / (p_{fit} - p_{mean})}{SS(fit) / (n - p_{fit})}$$

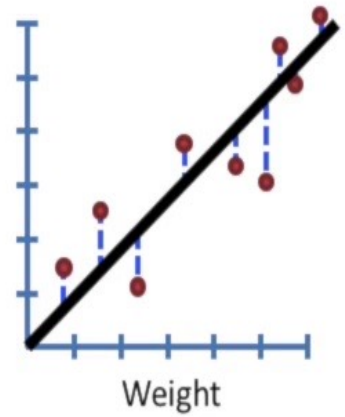




Linear Regression

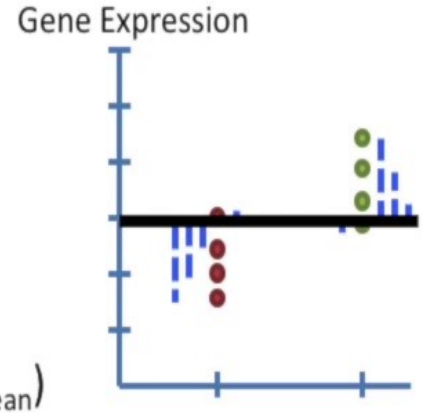
t-test

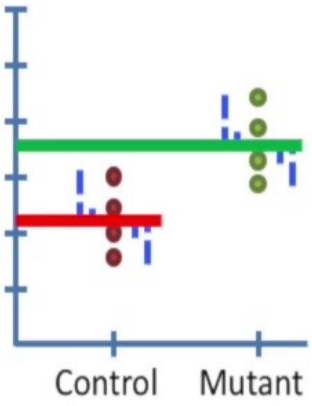




BAM!!!

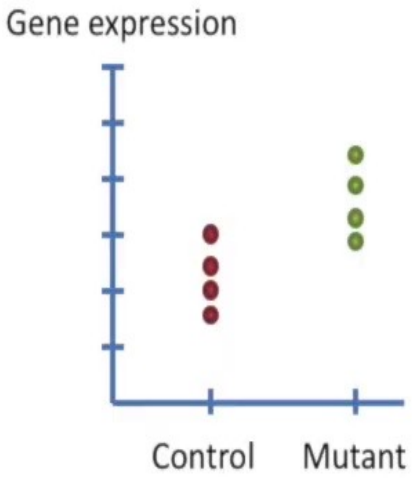
$$F = \frac{SS(mean) - SS(fit) / (p_{fit} - p_{mean})}{SS(fit) / (n - p_{fit})}$$





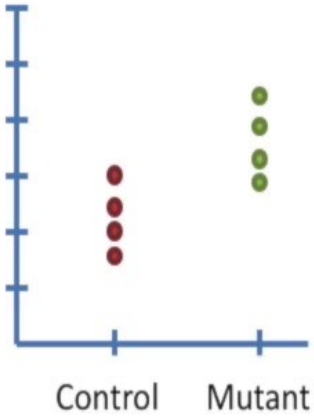
Let’s review what we’ve done so far..

The original data.

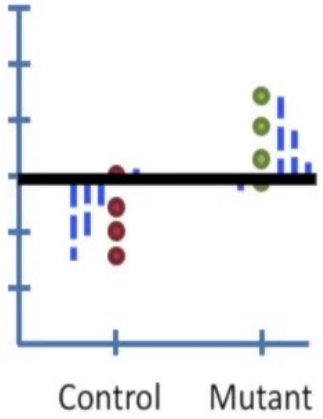


The original data.

Gene expression

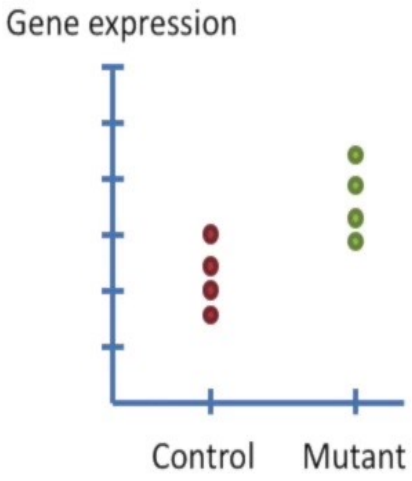


Calculate SS(mean)

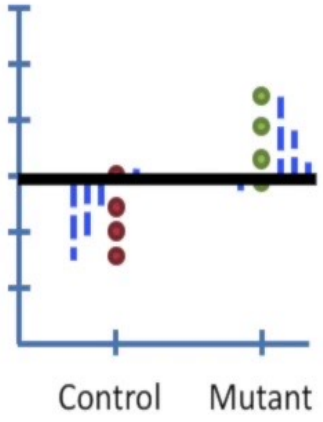


y = overall mean

The original data.

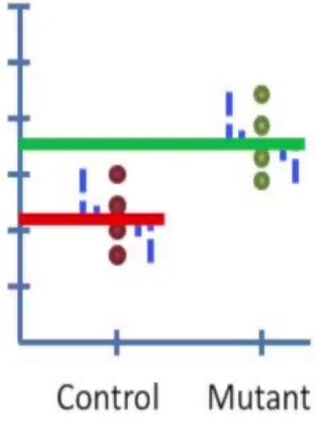


Calculate SS(mean)



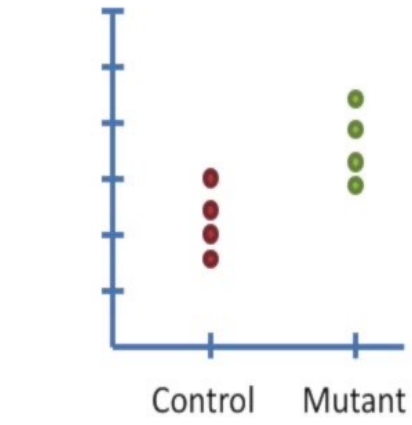
y = overall mean

Calculate SS(fit)

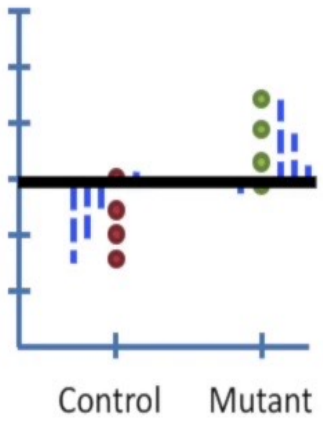


$y = \mathbf{mean}_{\mathbf{control}} + \mathbf{mean}_{\mathbf{mutant}}$

The original data.

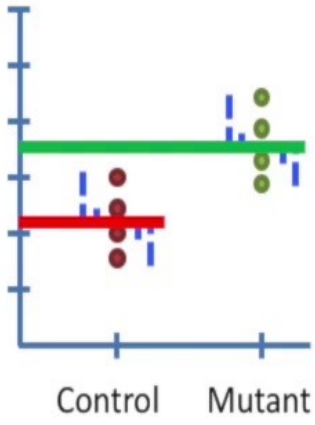


Calculate SS(mean)



y = overall mean

Calculate SS(fit)

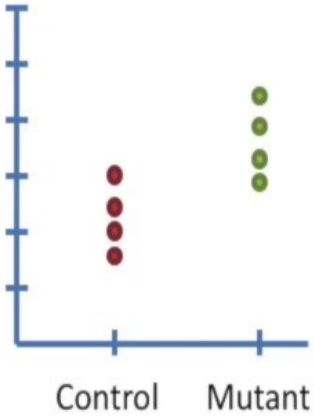


$y = \text{mean}_{\text{control}} + \text{mean}_{\text{mutant}}$

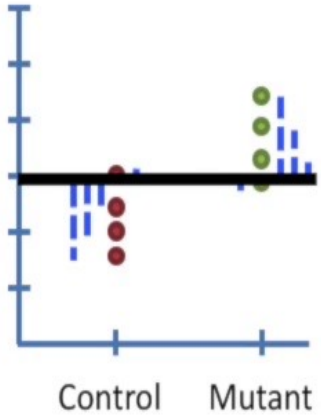
$$F = \frac{SS(\text{mean}) - SS(\text{fit}) \,/\, (p_{\text{fit}} - p_{\text{mean}})}{SS(\text{fit}) \,/\, (n - p_{\text{fit}})}$$

The original data.

Gene expression

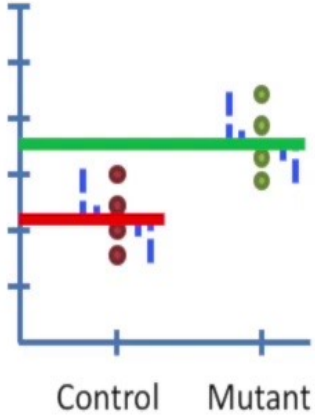


Calculate SS(mean)



y = overall mean

Calculate SS(fit)

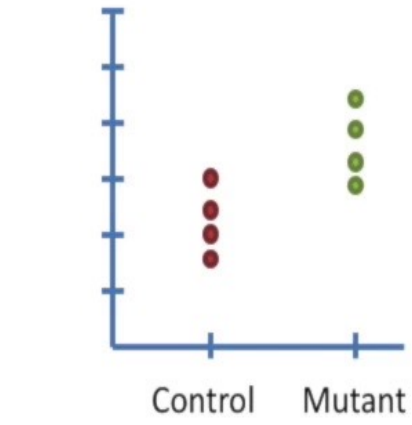


$y = \text{mean}_{\text{control}} + \text{mean}_{\text{mutant}}$

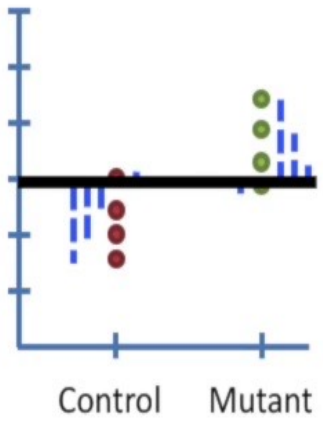
$p_{mean} = 1$

$$F = \frac{SS(mean) - SS(fit) / (p_{fit} - p_{mean})}{SS(fit) / (n - p_{fit})}$$

The original data.

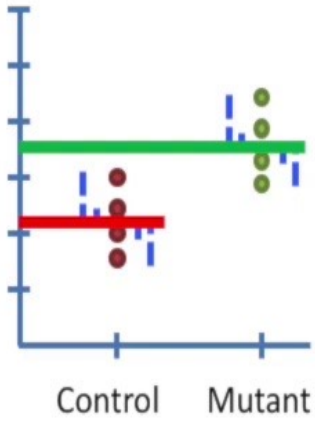


Calculate SS(mean)



y = overall mean

Calculate SS(fit)

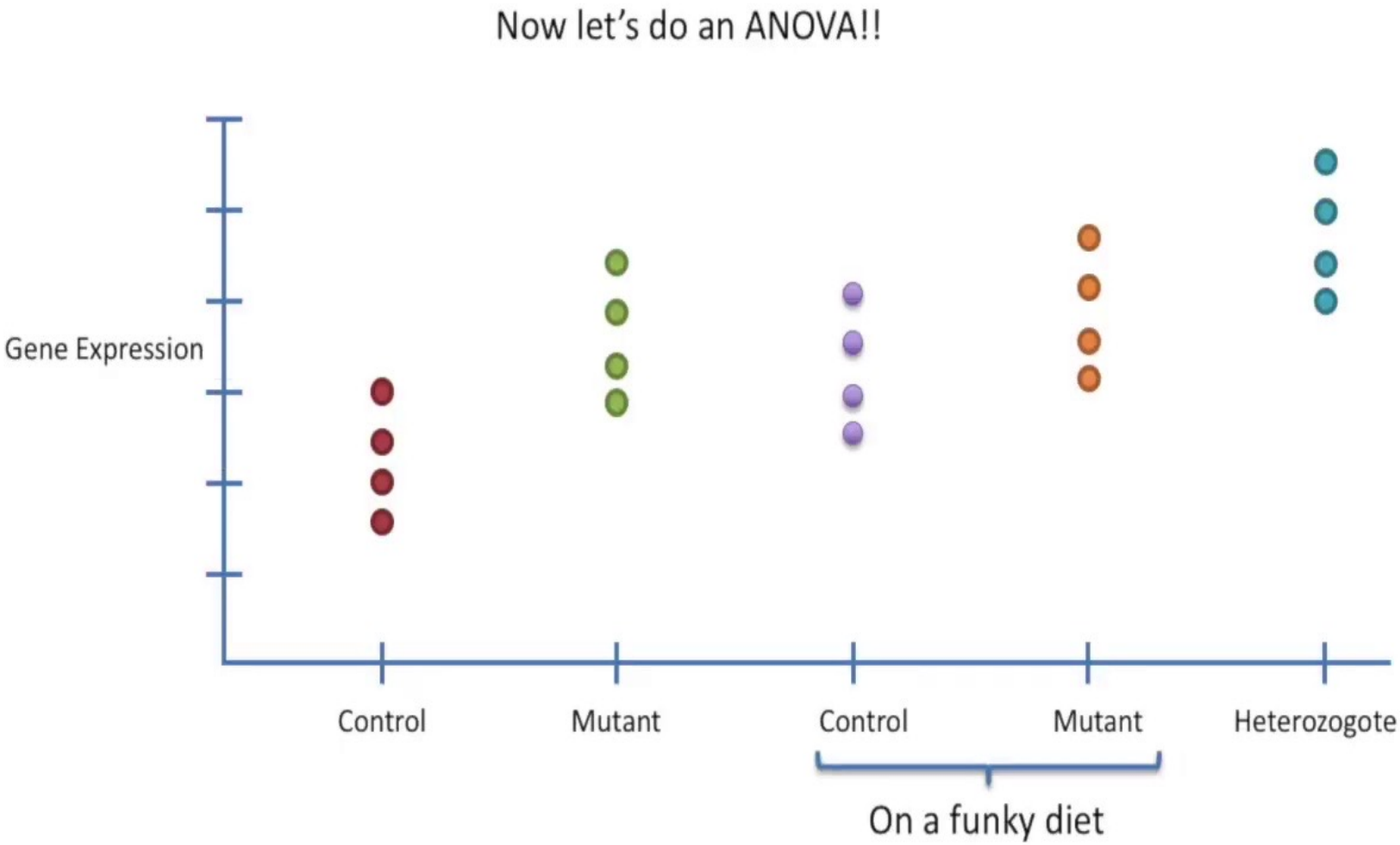


$y = \text{mean}_{\text{control}} + \text{mean}_{\text{mutant}}$

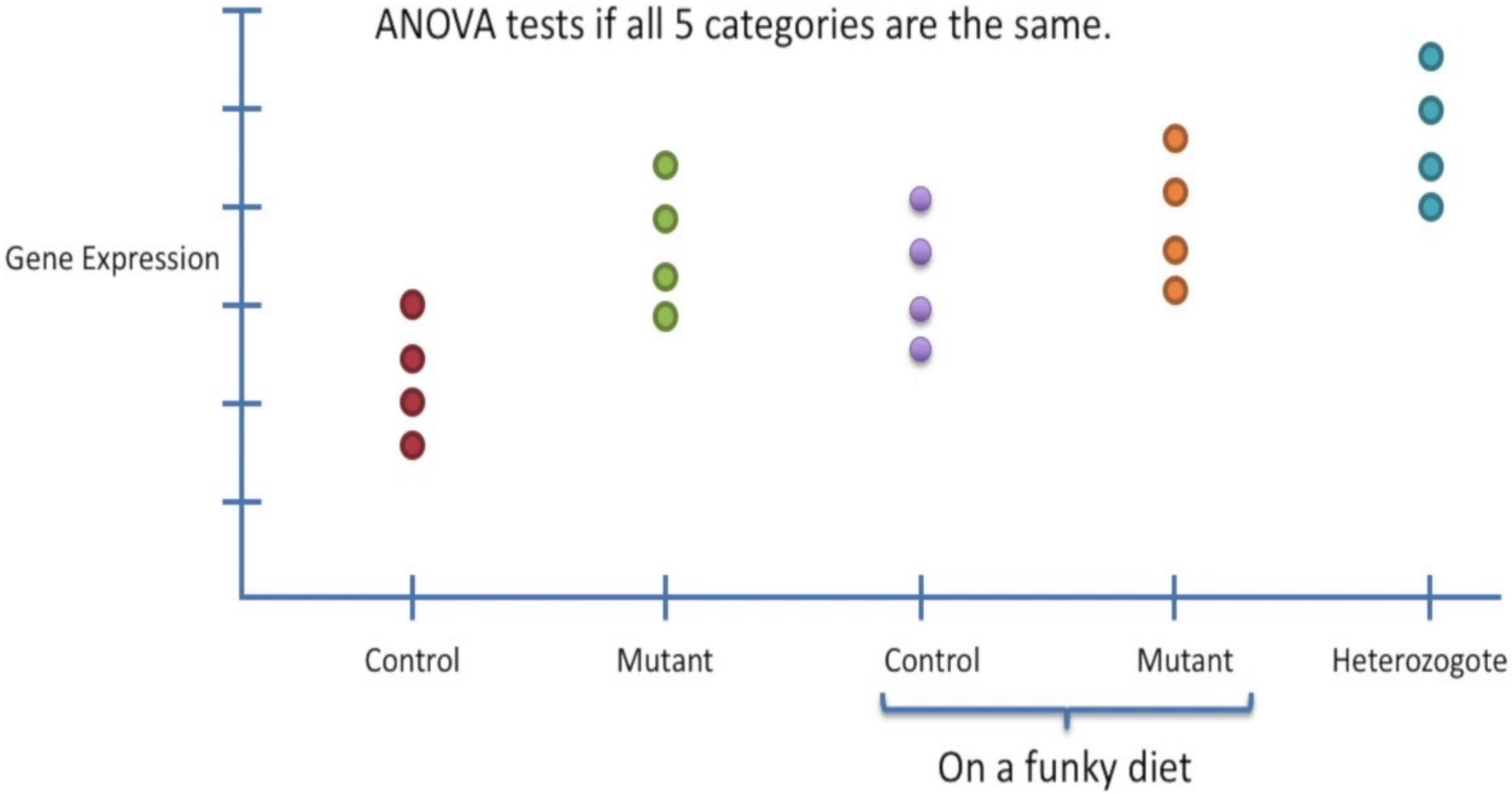
$$F = \frac{\text{SS(mean)} - \text{SS(fit)} \,/\, (p_{\text{fit}} - p_{\text{mean}})}{\text{SS(fit)} \,/\, (n - p_{\text{fit}})}$$

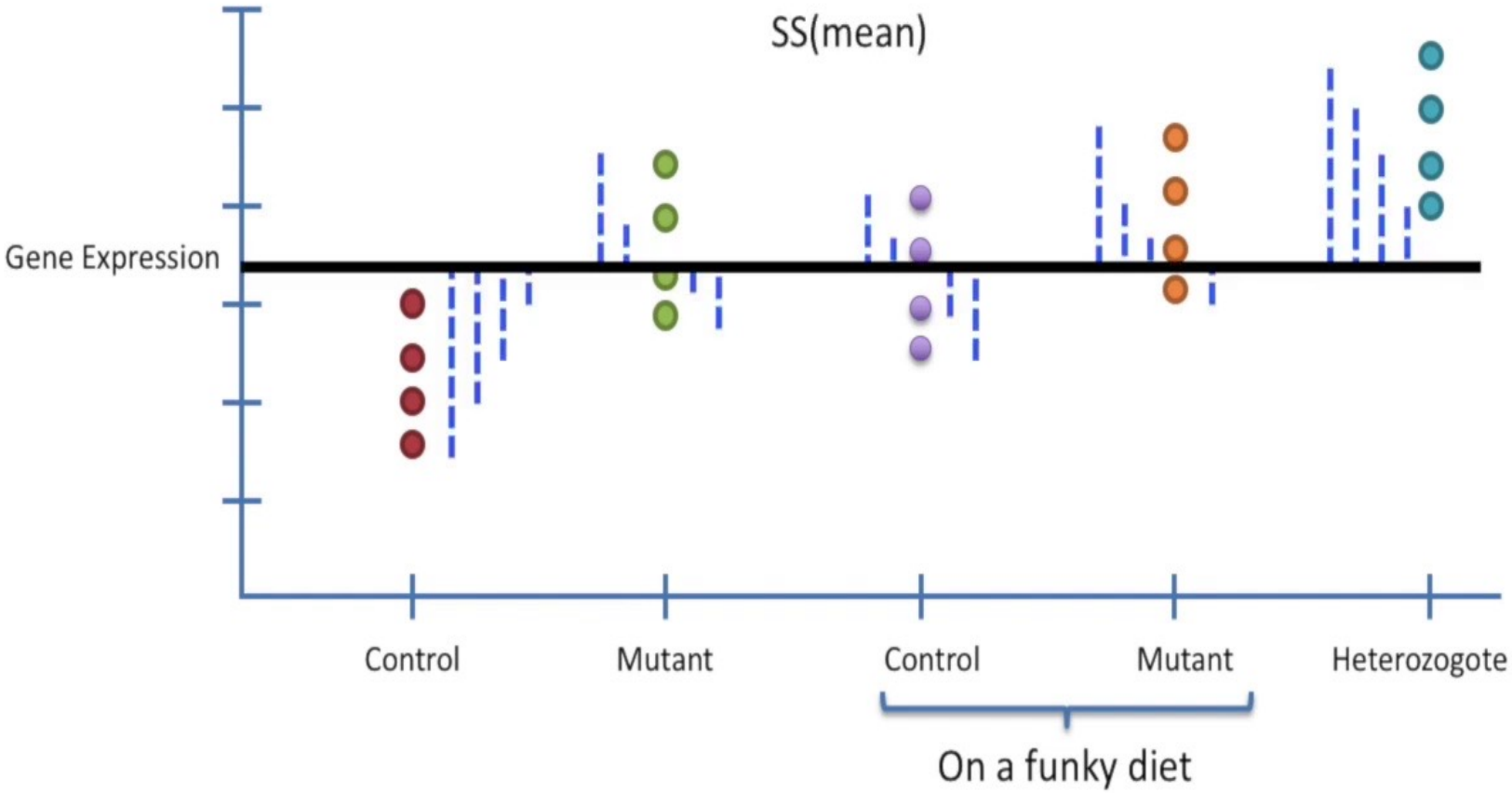
$p_{\text{mean}} = 1$

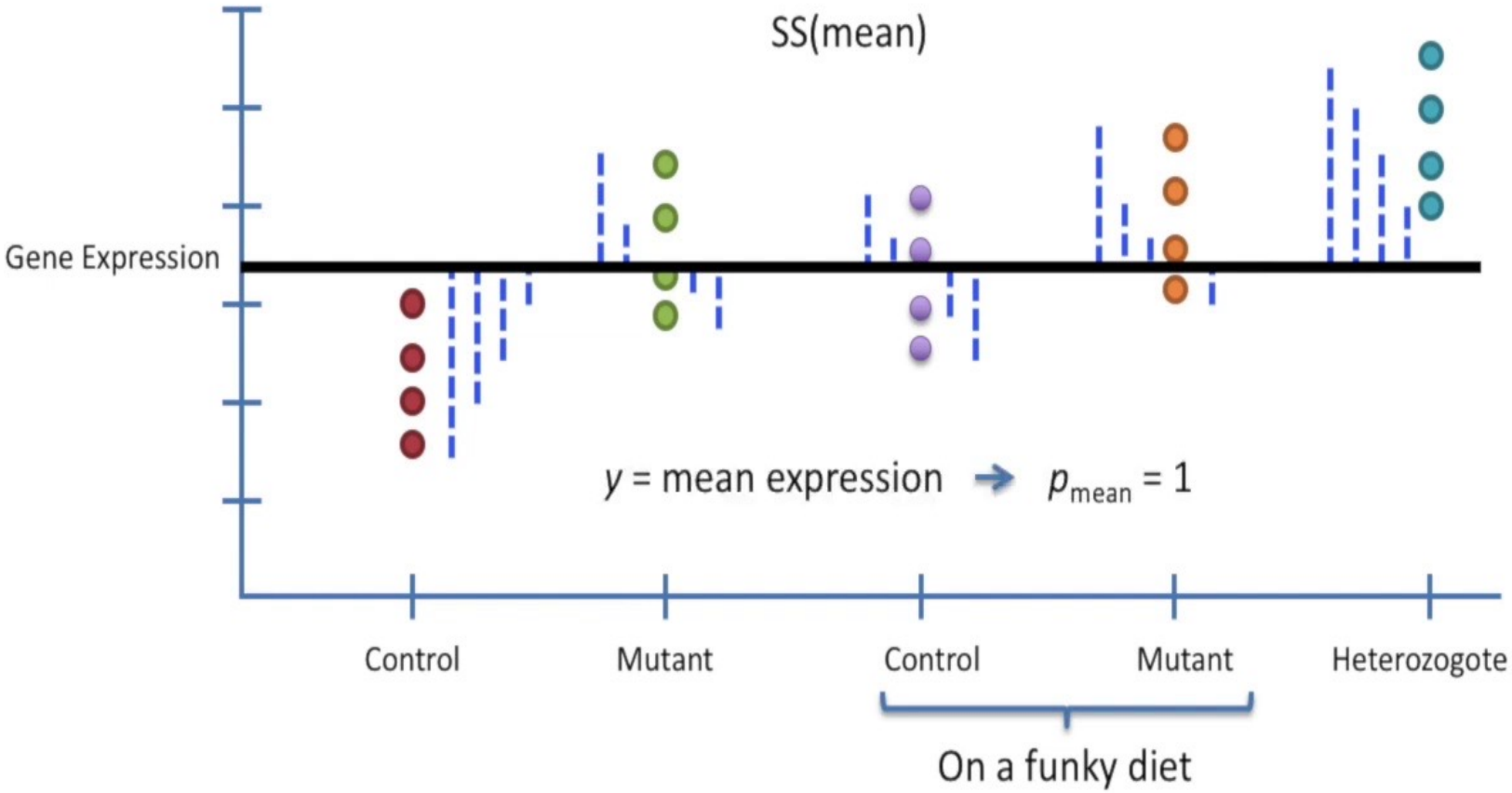
$p_{\text{fit}} = 2$

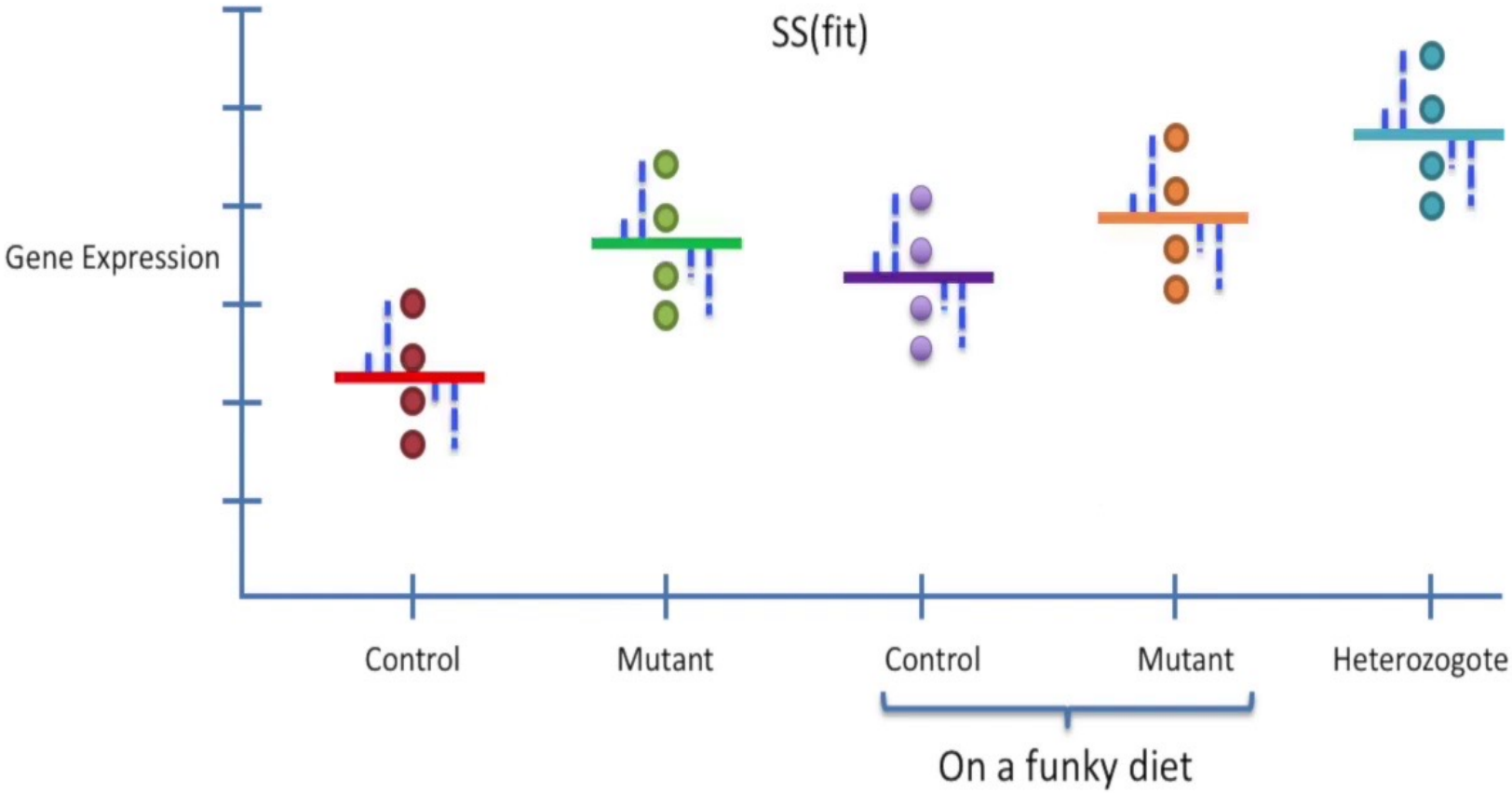

Now let's do an ANOVA!!
Gene Expression
Control
Mutant
Control
Mutant
Heterozogote
On a funky diet

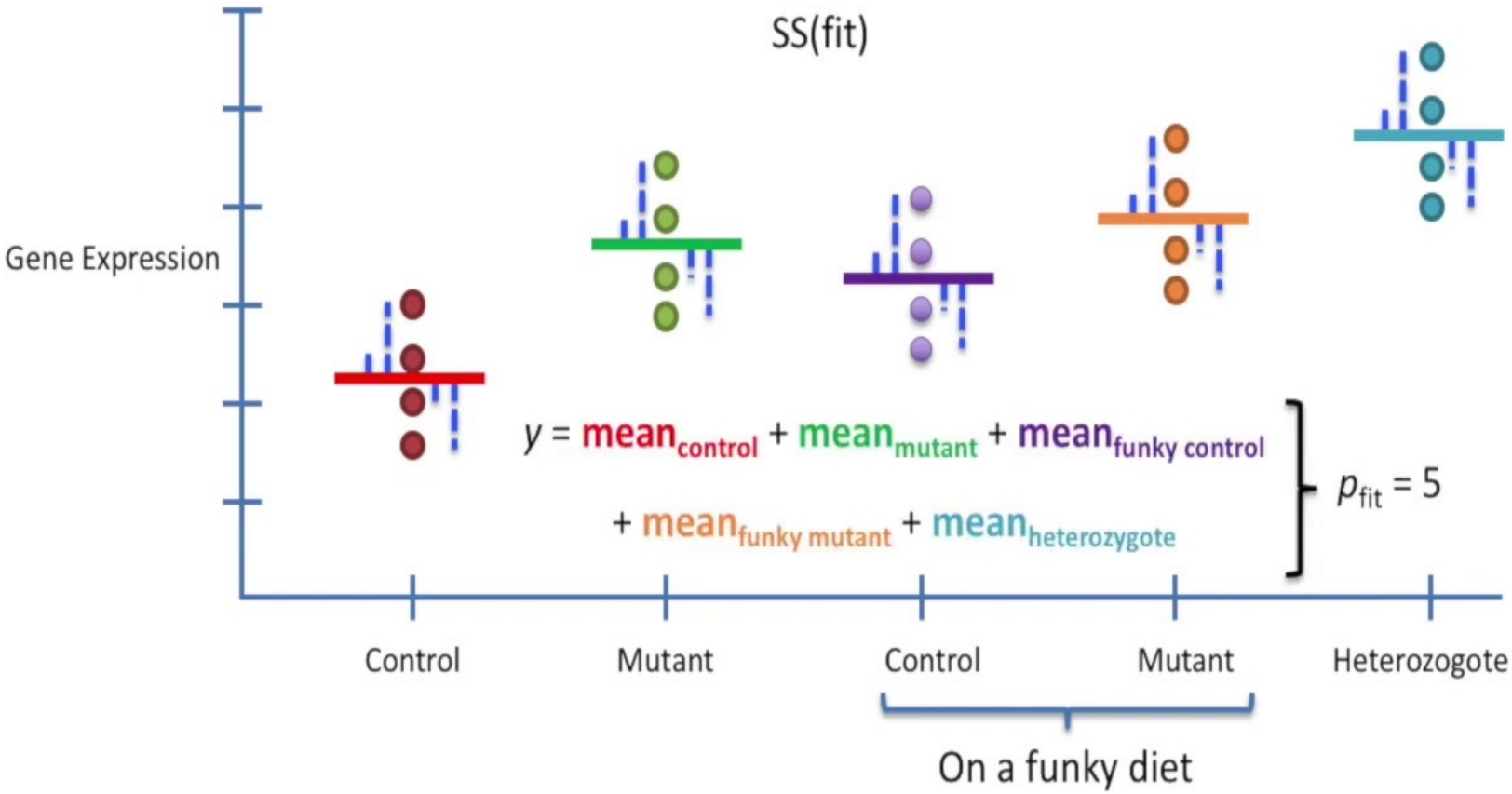
Now let's do an ANOVA!!




SS(mean)
Gene Expression
Control
Mutant
Control
Mutant
Heterozogote
On a funky diet

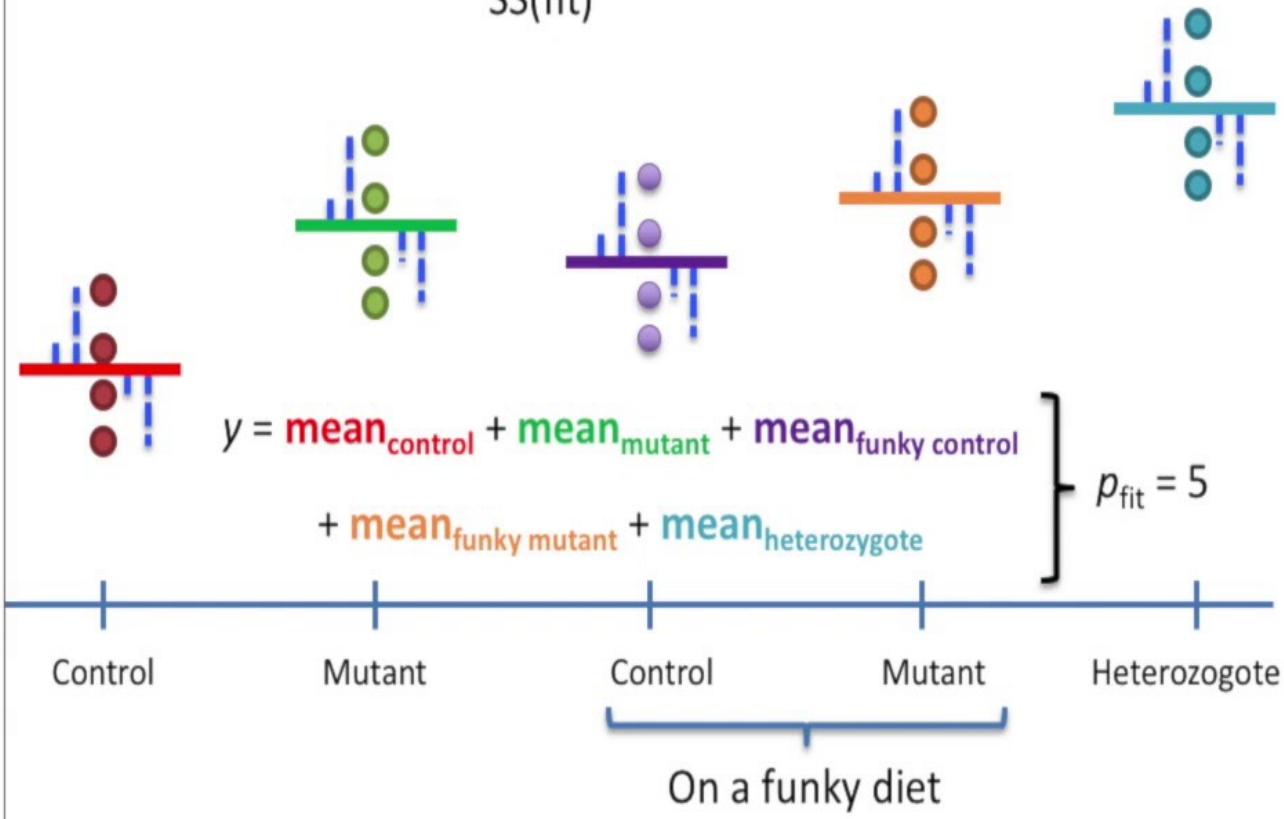

SS(mean)
Gene Expression
y = mean expression
p mean = 1
Control
Mutant
Control
Mutant
Heterozogote
On a funky diet

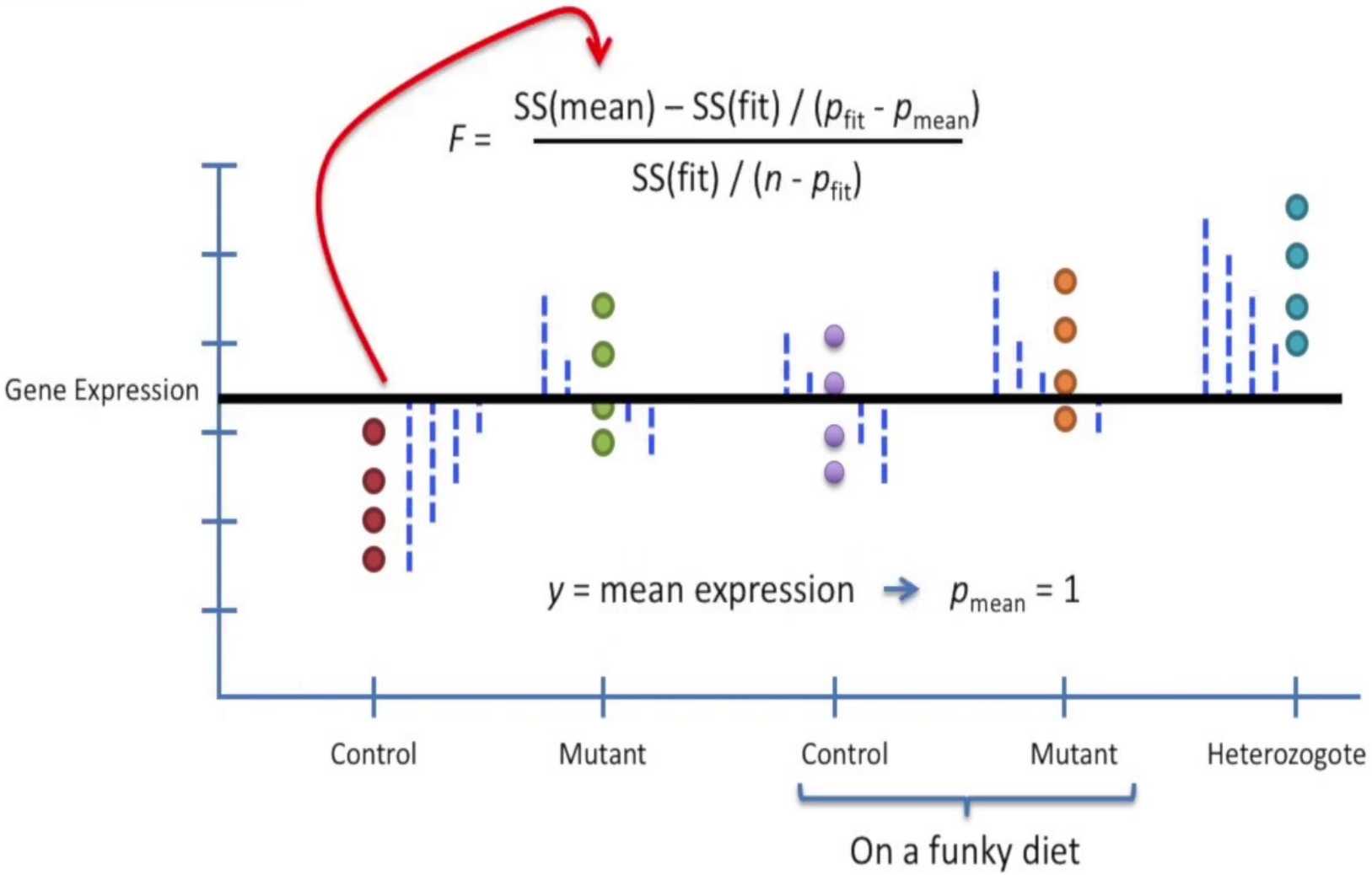

SS(fit)
Gene Expression
Control
Mutant
Control
Mutant
Heterozogote
On a funky diet

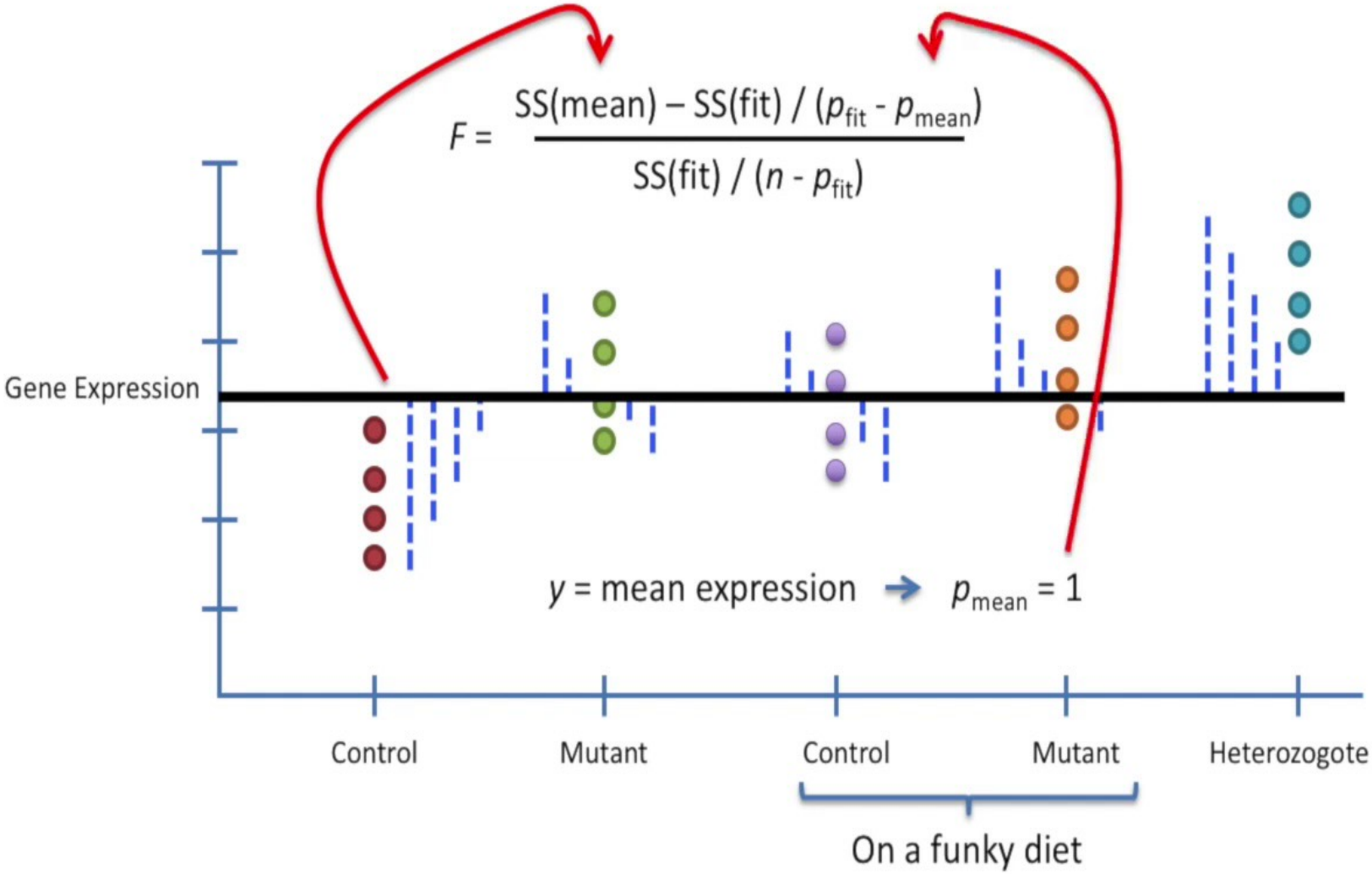

SS(fit)
Gene Expression
y = mean control + mean mutant + mean funky control
+ mean funky mutant + mean heterozygote
p fit = 5
Control
Mutant
Control
Mutant
Heterozogote
On a funky diet

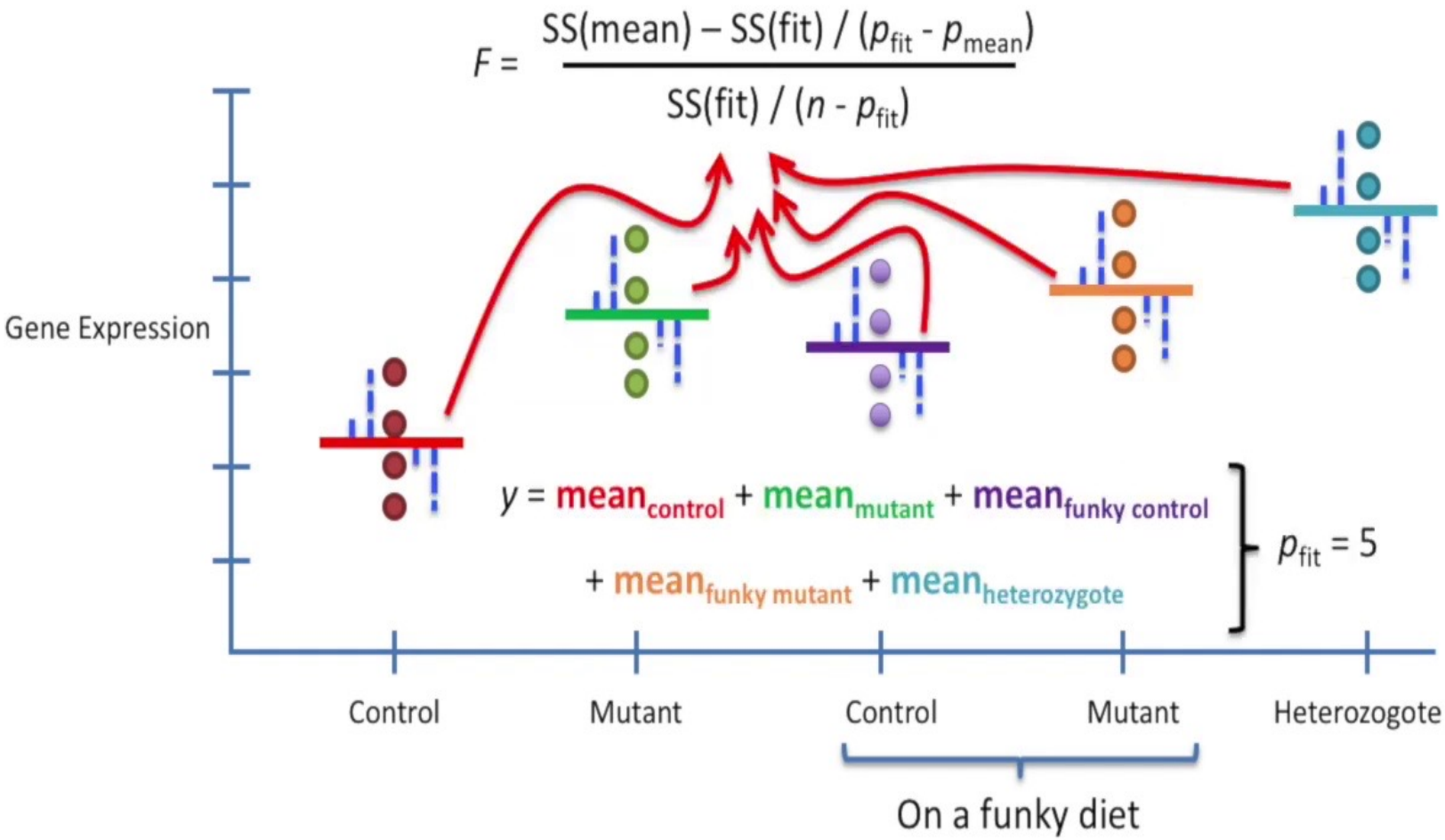
1	0	0	0	0
1	0	0	0	0
1	0	0	0	0
1	0	0	0	0
0	1	0	0	0
0	1	0	0	0
0	1	0	0	0
0	1	0	0	0
0	0	1	0	0
0	0	1	0	0
0	0	1	0	0
0	0	1	0	0
0	0	0	1	0
0	0	0	1	0
0	0	0	1	0
0	0	0	1	0
0	0	0	0	1
0	0	0	0	1
0	0	0	0	1
0	0	0	0	1

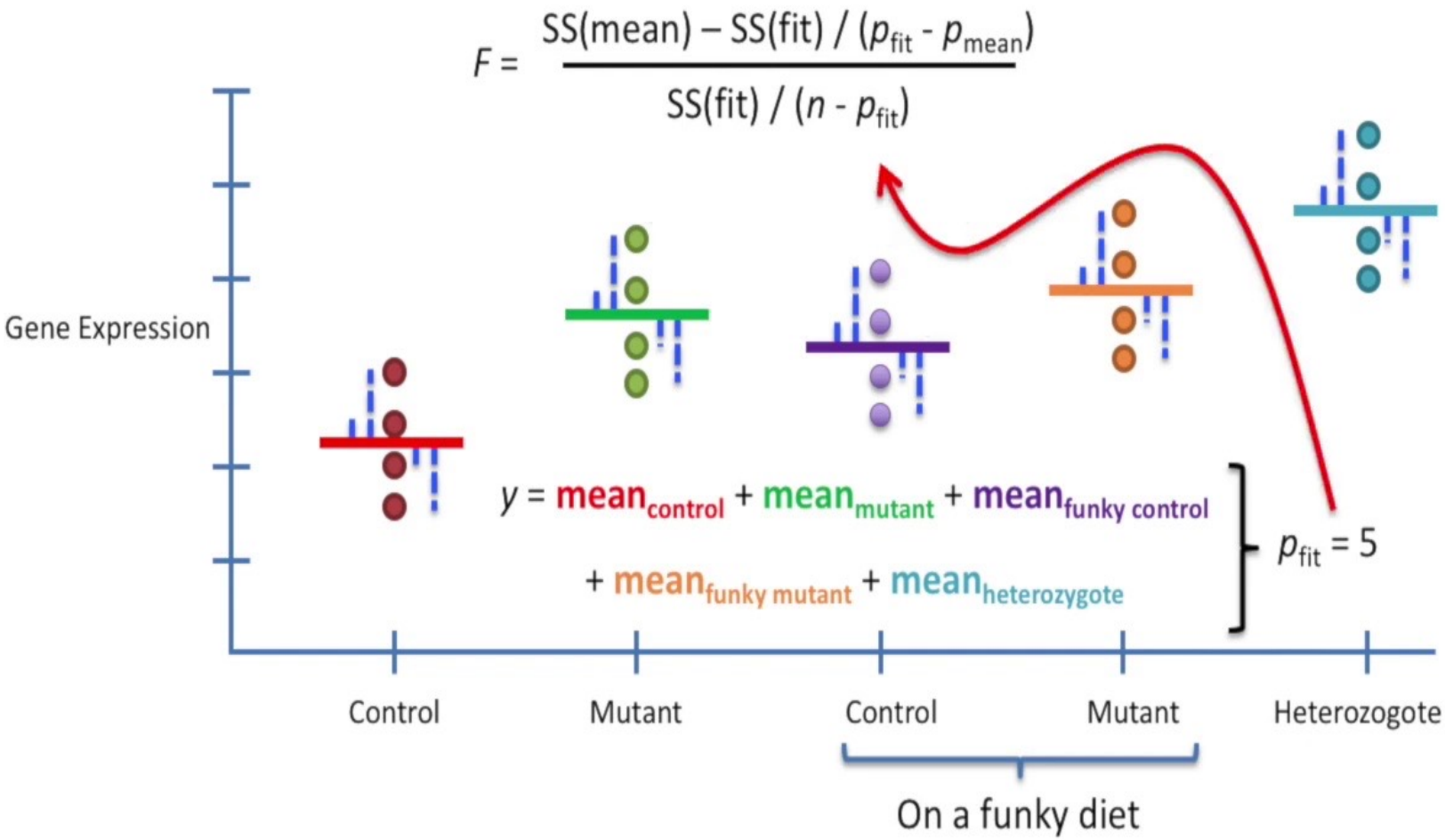
Here's what the design matrix looks like - one column per category.




$F = \frac{SS(mean) - SS(fit) / (p_{fit} - p_{mean})}{SS(fit) / (n - p_{fit})}$
Gene Expression
y = mean expression → $p_{mean} = 1$
Control
Mutant
Control
Mutant
Heterozogote
On a funky diet

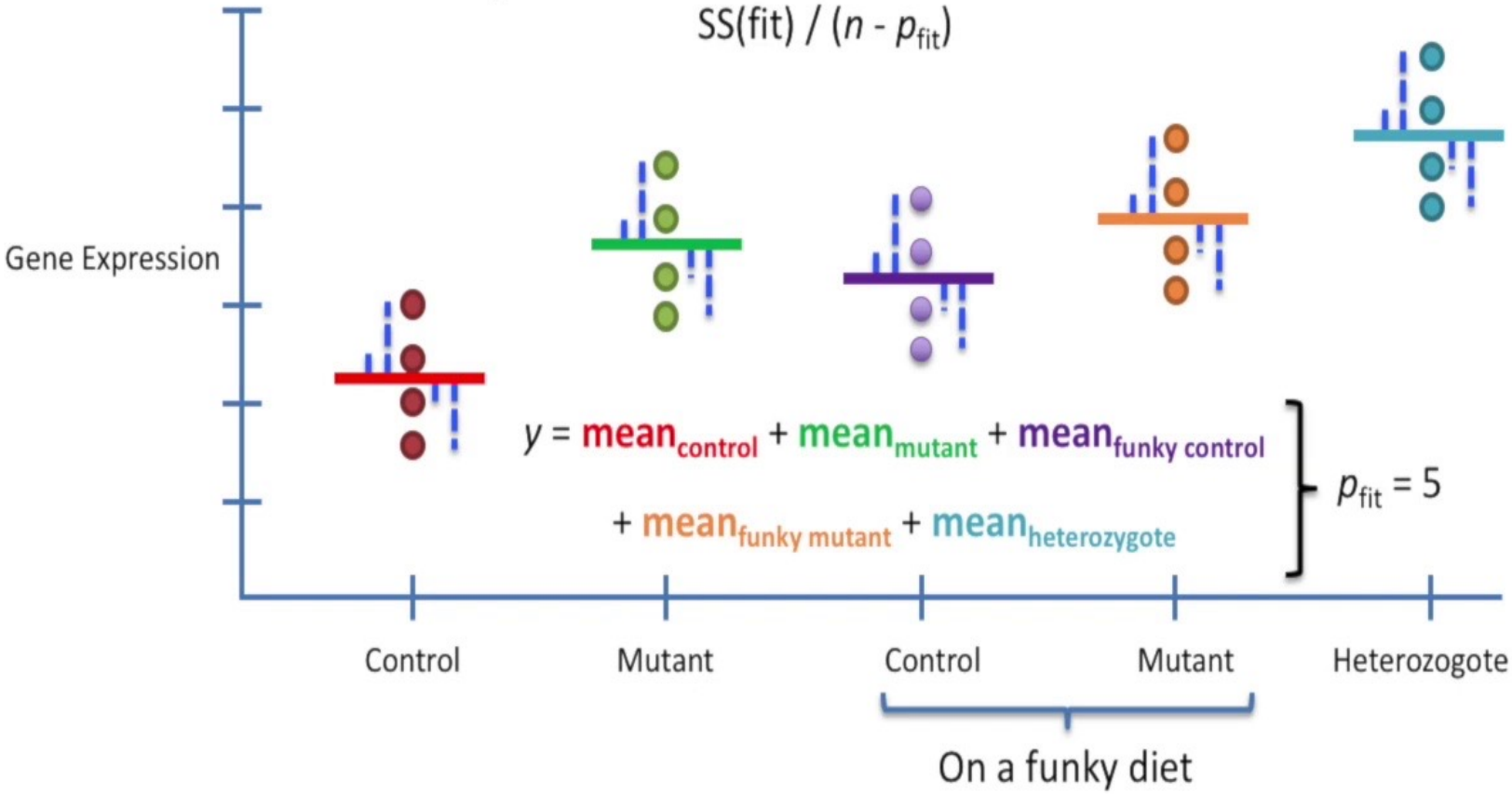

$F = \frac{SS(mean) - SS(fit) / (p_{fit} - p_{mean})}{SS(fit) / (n - p_{fit})}$
Gene Expression
y = mean expression → $p_{mean} = 1$
Control
Mutant
Control
Mutant
Heterozogote
On a funky diet


$F = \frac{SS(mean) - SS(fit) / (p_{fit} - p_{mean})}{SS(fit) / (n - p_{fit})}$
Gene Expression
$y = mean_{control} + mean_{mutant} + mean_{funky\ control}$
$+ mean_{funky\ mutant} + mean_{heterozygote}$
$p_{fit} = 5$
Control
Mutant
Control
Mutant
Heterozogote
On a funky diet


$F = \frac{SS(mean) - SS(fit) / (p_{fit} - p_{mean})}{SS(fit) / (n - p_{fit})}$
Gene Expression
$y = mean_{control} + mean_{mutant} + mean_{funky\ control}$
$+ mean_{funky\ mutant} + mean_{heterozygote}$
$p_{fit} = 5$
Control
Mutant
Control
Mutant
Heterozogote
On a funky diet

$$F = \frac{SS(mean) - SS(fit) / (p_{fit} - p_{mean})}{SS(fit) / (n - p_{fit})}$$

TRIPLE BAM!!!!



One last important detail before we're done...

The design matrices that I've shown you are not the standard design matrices used for doing t-tests and ANOVA.

One last important detail before we're done...

1	0
1	0
1	0
1	0
0	1
0	1
0	1
0	1

The design matrices that I've shown you are not the standard design matrices used for doing t-tests and ANOVA.

This is what we used for the t-test in this StatQuest...

$y = \mathbf{mean}_{\mathbf{control}} + \mathbf{mean}_{\mathbf{mutant}}$

One last important detail before we're done...

1	0
1	0
1	0
1	0
0	1
0	1
0	1
0	1

The design matrices that I've shown you are not the standard design matrices used for doing t-tests and ANOVA.

This is what we used for the t-test in this StatQuest...

$y = \text{mean}_{\text{control}} + \text{mean}_{\text{mutant}}$

...but this is a more common design matrix. We'll talk about this one and other, more elaborate designs in the next StatQuest.

1	0
1	0
1	0
1	0
1	1
1	1
1	1
1	1

The End!!!!