BIO-5023YB
2020
Spring term – week 3
Linear models continued

Dr Philip Leftwich – p.leftwich@uea.ac.uk

Learning outcomes

- Understand the Ordinary Least Squares method of regression

- Calculate F and t for hypothesis testing

- Practice results writing

Recap on Ordinary Least Squares

Recall that we are using ________ models to quantify the variability in our datasets.

This fits a _______ line equation which produces the ________ squares difference between our slope of the line and the data points.

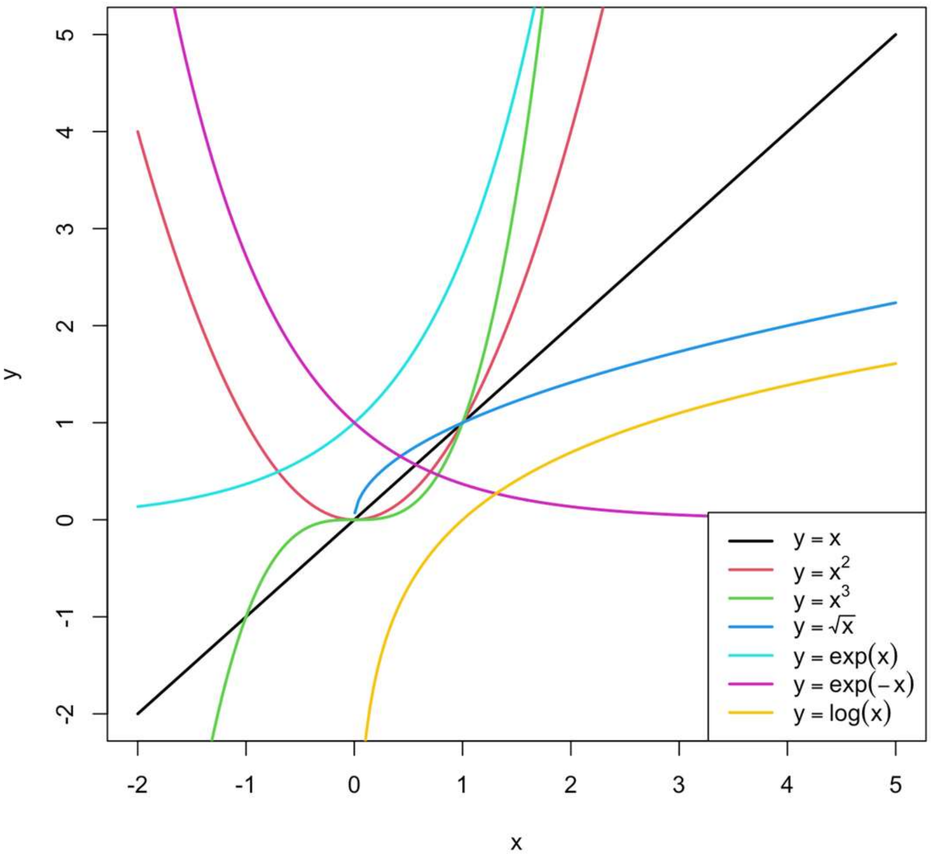
Once we have fit a model to our data we can describe it in terms of two parts

__________ - What is explained by the mode

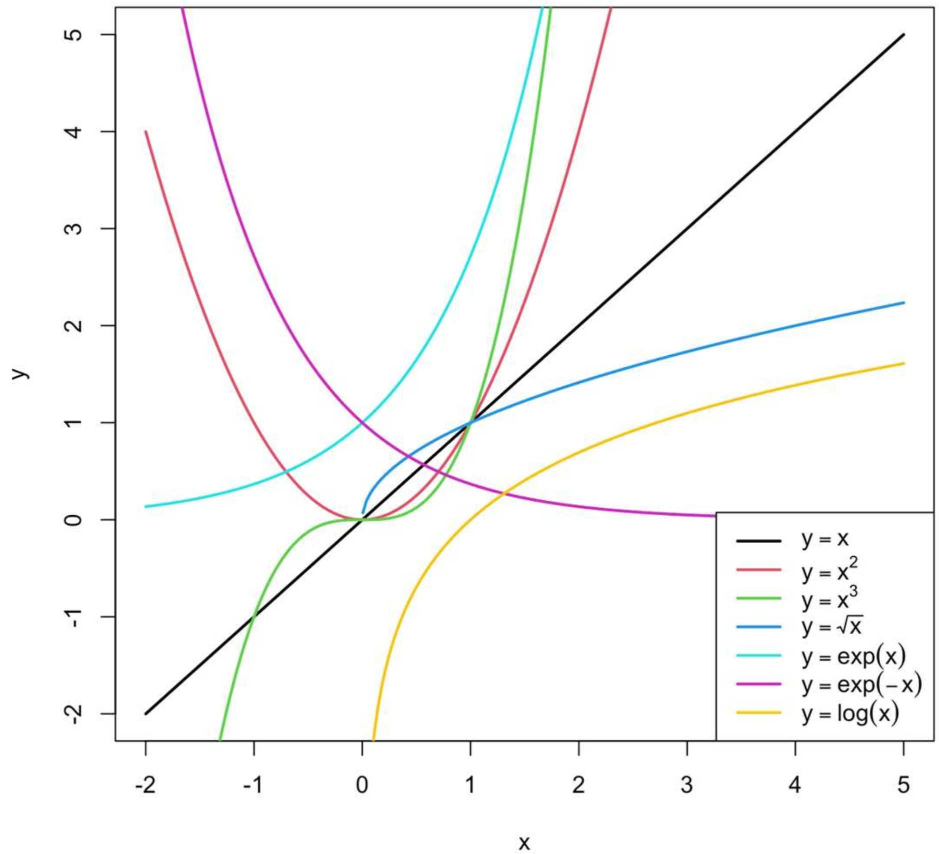
___________ - The variance which is not explained by the model

What is the first assumption we met when using linear models?

What is the first assumption we met when using linear models?



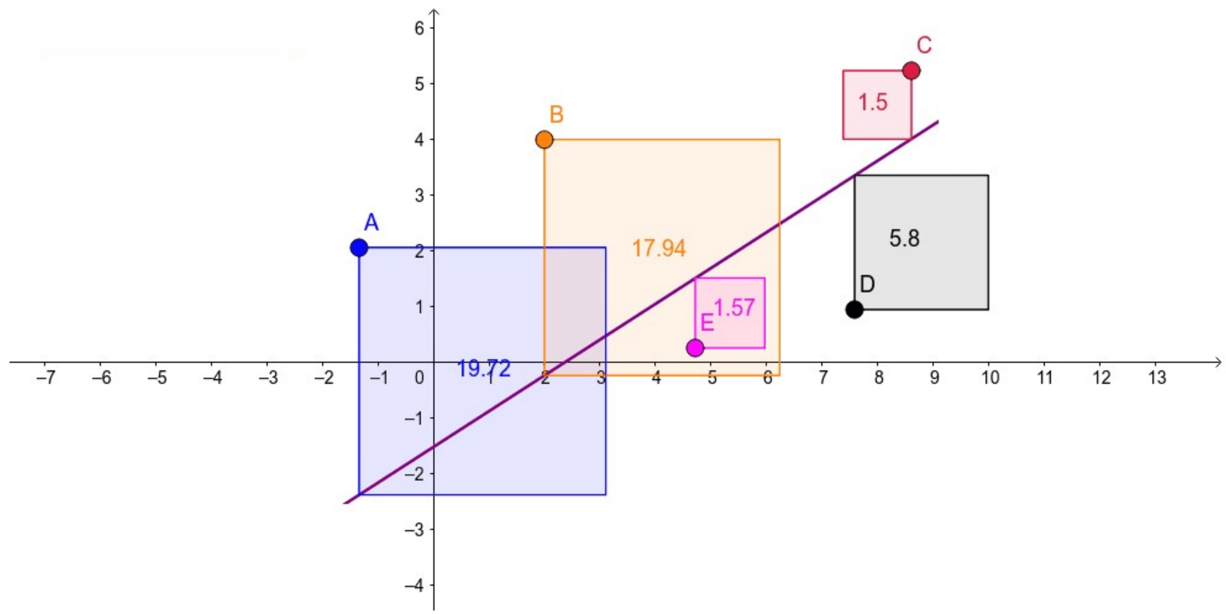
What is the first assumption we met when using linear models?



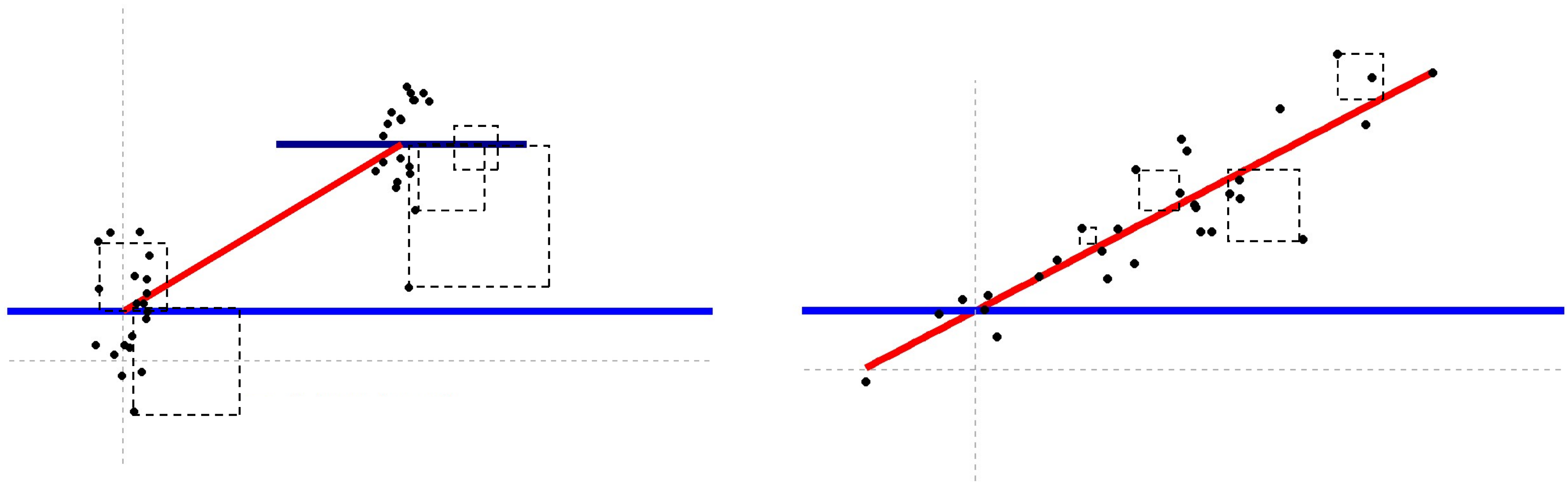
Later we will encounter methods (including data transformation) that often allow us to “approximate” a linear relationship

Recap on Ordinary Least Squares

OLS ~ Draws the regression line in the way that produces the smallest value of the *squared* residuals



Recap on Ordinary Least Squares



This least squares method works just as well for fitting a line to compare two (or more) means as it does to fit a regression

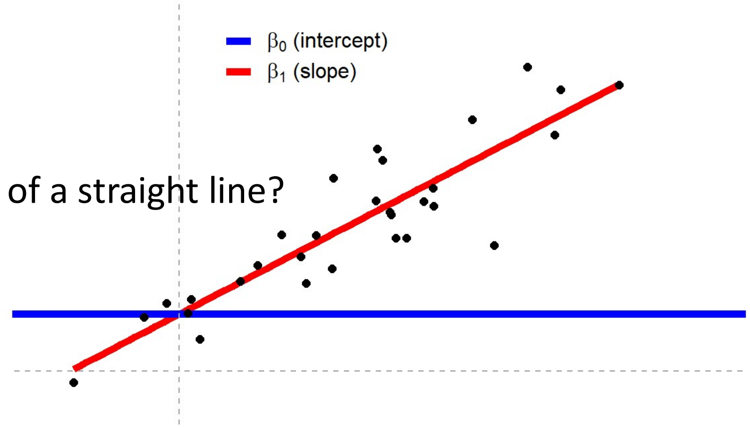
Equation for the straight line

We use OLS to fit the line but what is the equation that explains the fit of a straight line?

Equation for the straight line

We use OLS to fit the line but what is the equation that explains the fit of a straight line?

$y = mx + c$

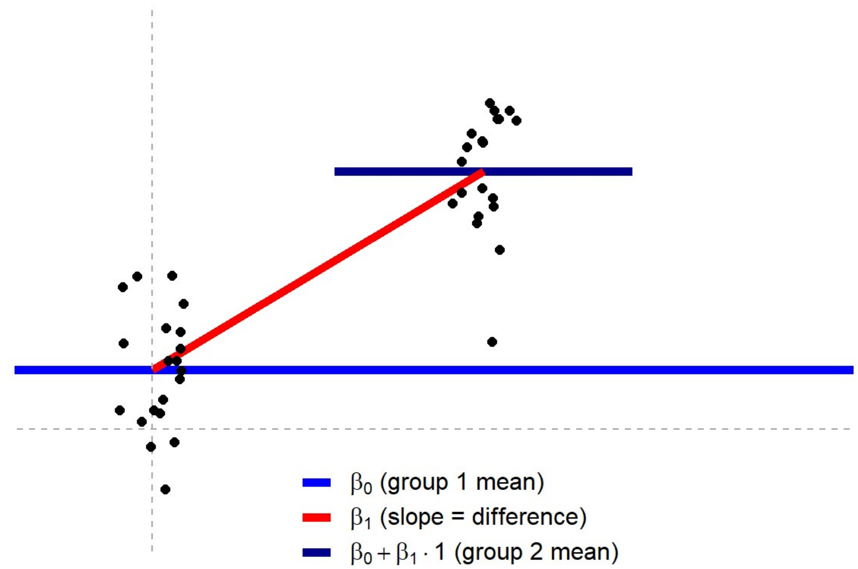


When written to describe a general linear model

You may see this as

$y = \beta 0 + \beta 1 * x$

This is <u>exactly</u> the same equation just shuffled round



Equation for the straight line

We use OLS to fit the line but what is the equation that explains the fit of a straight line?

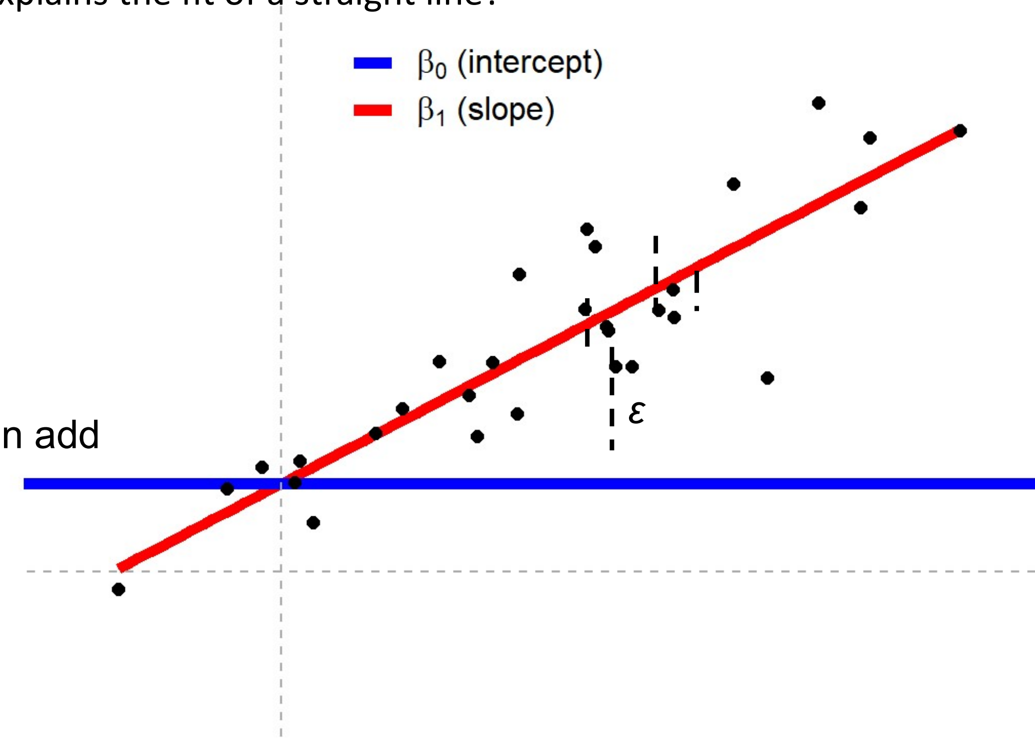
*y = β0 + β1 * x*

Describes the <u>fit</u> of the model

This is the bit we care about.

To produce the <u>full</u> equation for a linear model we can add a term for the <u>residuals</u>

$y = \beta 0 + \beta 1x + \varepsilon$

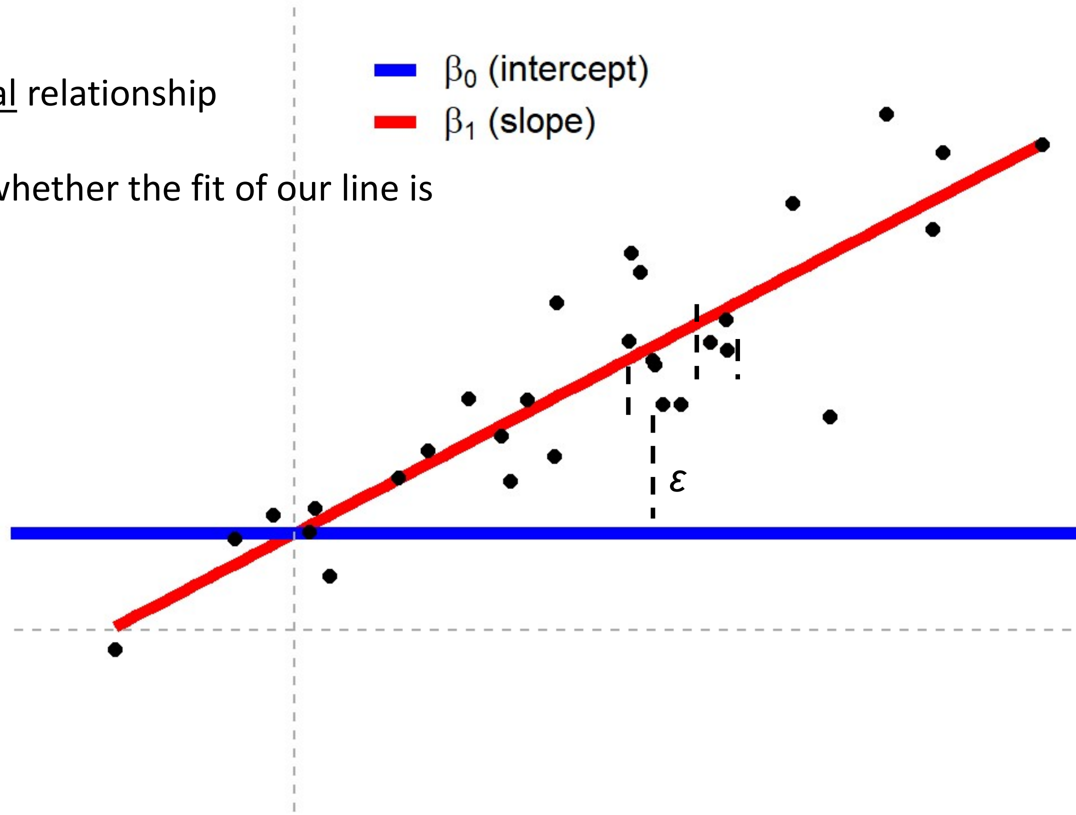


Equation for the straight line

The fit of our model describes the nature of our biological relationship

The comparison of our fit and the residuals determines whether the fit of our line is significantly different from the intercept

Remember our Null Hypothesis of *no difference* or *no relationship* would plot a flat line



Hypothesis testing

We typically care whether our relationship/difference is *significant.*

We have already seen how an understanding of a linear model gives us *more* information about the nature of our relationship/ difference than the traditional ANOVA approach than just *significance*.

```
Call:
lm(formula = height ~ type, data = darwin)

Residuals:
    Min      1Q  Median      3Q     Max 
-8.1917 -1.0729  0.8042  1.9021  3.3083 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  20.1917     0.7592  26.596   <2e-16 ***
typeSelf     -2.6167     1.0737  -2.437   0.0214 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.94 on 28 degrees of freedom
Multiple R-squared:  0.175,     Adjusted R-squared:  0.1455 
F-statistic:  5.94 on 1 and 28 DF,  p-value: 0.02141
```

Our maize plant data from the last workshop

Hypothesis testing

We typically care whether our relationship/difference is *significant.*

We have already seen how an understanding of a linear model gives us *more* information about the nature of our relationship/ difference than the traditional ANOVA approach than just *significance*.

```
Call:
lm(formula = height ~ type, data = darwin)

Residuals:
    Min      1Q  Median      3Q     Max 
-8.1917 -1.0729  0.8042  1.9021  3.3083 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  20.1917     0.7592  26.596   <2e-16 ***
typeSelf     -2.6167     1.0737  -2.437   0.0214 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.94 on 28 degrees of freedom
Multiple R-squared:  0.175,	Adjusted R-squared:  0.1455 
F-statistic:  5.94 on 1 and 28 DF,  p-value: 0.02141
```

Calculating *F* and *R-squared*

Calculating *F* & *R-squared* allow us to determine the amount of variance in our data that is explained by the fit of the model and then determine whether this is significantly bigger than the residual variance.

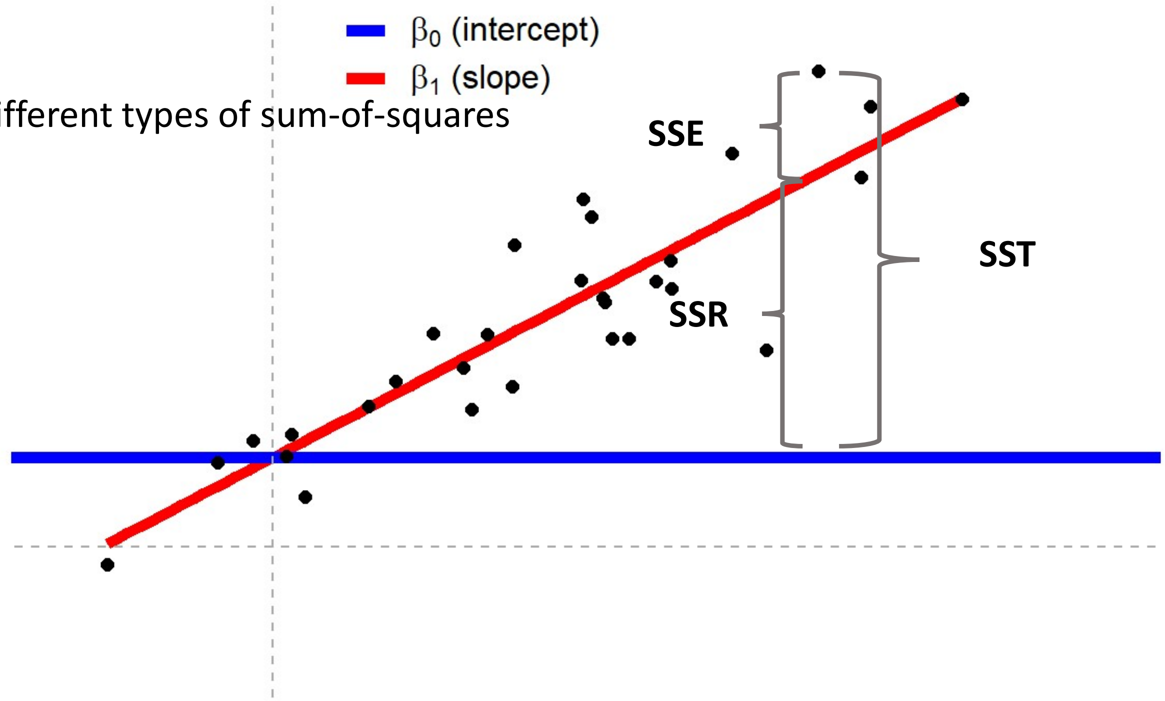
You may also have heard me refer to this as the *signal-to-noise* ratio.

Last lecture we were *briefly* introduced to our different types of sum-of-squares

SST – total sum of squares

SSR – sum of squares for the regression

SSE – sum of squares for error/residuals



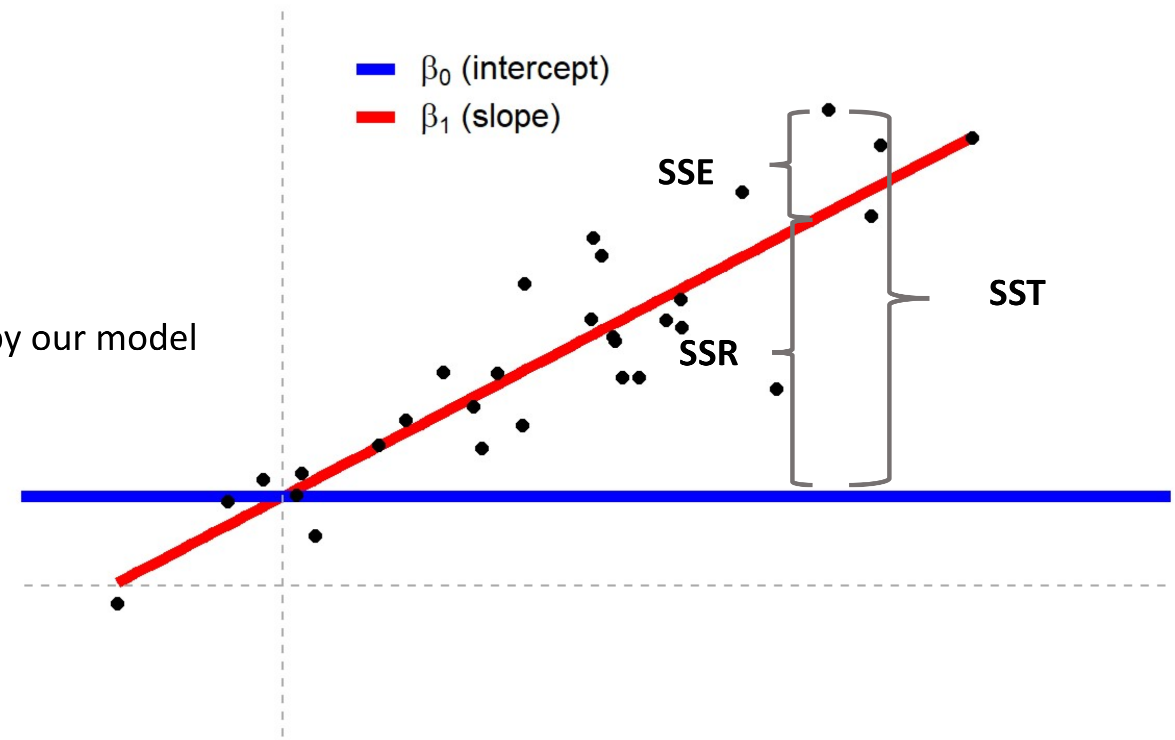
Calculating *F* and *R-squared*

SST – total sum of squares
SSR – sum of squares for the regression
SSE – sum of squares for error/residuals

$$R^2 = \frac{SSR}{SST}$$

A perfect fit would produce an R-squared of 1

e.g. 100% of our dataset variance is explained by our model



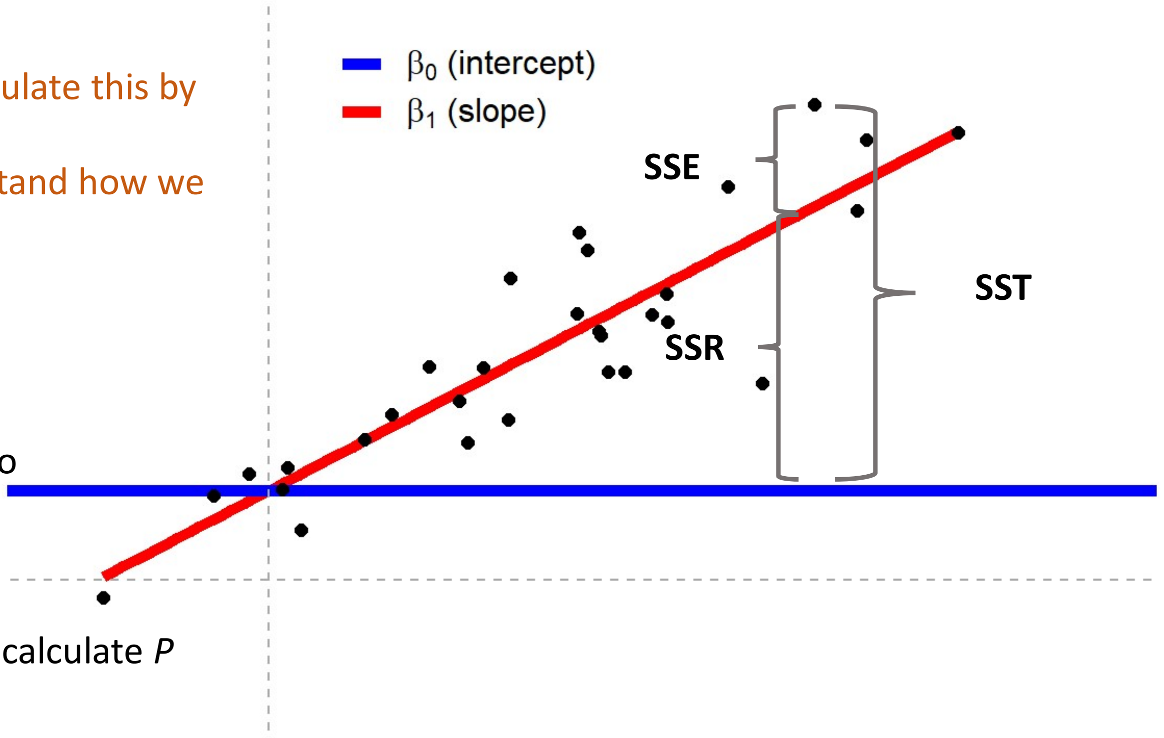
Calculating *F* and *R-squared*

SST – total sum of squares
SSR – sum of squares for the regression
SSE – sum of squares for error/residuals

$F = \frac{SSR\,/(n-p)}{SSE/(p-1)}$

We don't need to calculate this by hand *phew
But it helps us understand how we calculate *P*

n = *sample size*
p = number of treatments



The larger F is, the larger our signal-to-noise ratio

Report as $F_{p,n\text{-}p}$ =

Our *F-value* with the sample size can be used to calculate *P*

F vs *t*

```
Call:
lm(formula = height ~ type, data = darwin)

Residuals:
    Min      1Q  Median      3Q     Max 
-8.1917 -1.0729  0.8042  1.9021  3.3083 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  20.1917     0.7592  26.596   <2e-16 ***
typeSelf     -2.6167     1.0737  -2.437   0.0214 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.94 on 28 degrees of freedom
Multiple R-squared:  0.175,	Adjusted R-squared:  0.1455 
F-statistic:  5.94 on 1 and 28 DF,  p-value: 0.02141
```

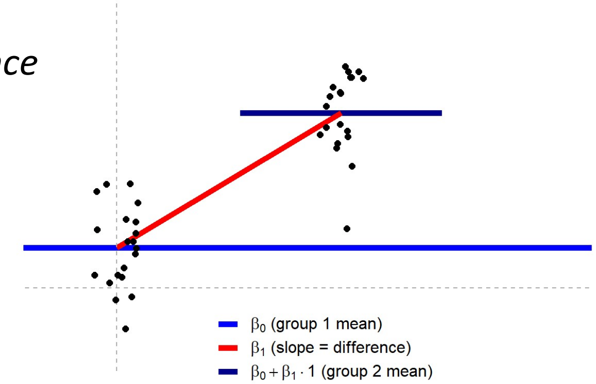
?

F vs *t*

t is calculated from the *estimate* divided by the *standard error of the difference*

For a difference model this is the *difference between the two means*

For a regression it's the *slope* both are found on the model summary as the <u>estimate</u>



$$t = \frac{-2.62}{1.07}$$

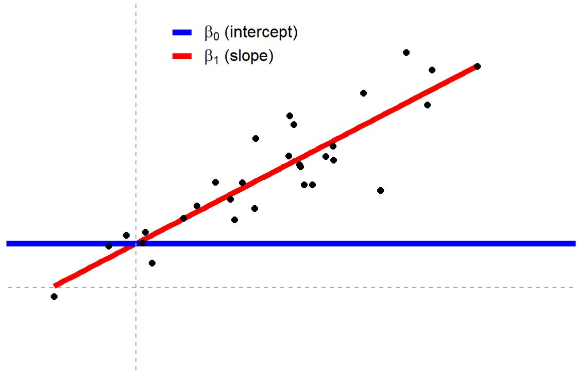
$$t = 2.44$$

```
Call:
lm(formula = height ~ type, data = darwin)

Residuals:
    Min      1Q  Median      3Q     Max 
-8.1917 -1.0729  0.8042  1.9021  3.3083 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  20.1917     0.7592  26.596   <2e-16 ***
typeSelf     -2.6167     1.0737  -2.437   0.0214 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.94 on 28 degrees of freedom
Multiple R-squared:  0.175,	Adjusted R-squared:  0.1455 
F-statistic:  5.94 on 1 and 28 DF,  p-value: 0.02141
```



F vs *t*

So how does *t* compare to *F*

In a simple model they are the same:

- *t* is calculated directly from the means and errors in our model

- *F* is calculated from the squared errors in our model

So is it possible that $t^2 = F$

$-2.437^2 = 5.94$


```
Call:
lm(formula = height ~ type, data = darwin)

Residuals:
    Min      1Q  Median      3Q     Max 
-8.1917 -1.0729  0.8042  1.9021  3.3083 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  20.1917     0.7592  26.596   <2e-16 ***
typeSelf     -2.6167     1.0737  -2.437   0.0214 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.94 on 28 degrees of freedom
Multiple R-squared:  0.175,	Adjusted R-squared:  0.1455 
F-statistic:  5.94 on 1 and 28 DF,  p-value: 0.02141
```

Why F and t?

If in our example here *t* and *F* are both essentially the same. Why have both?

t can only be calculated for single predictors – it cannot be used for more than one at a time

F can be used no matter the number of different predictors in our model

In more complex models we will see that multiple *t-values* are generated to test the significance of each predictor separately within a model and *F* is used to test the significance of the *whole* model.

Bringing it all together!

There is a significant difference in the heights of our cross-bred and selfed maize plants (P<0.05)

Can we do any better than this?

```
Call:
lm(formula = height ~ type, data = darwin)

Residuals:
    Min      1Q  Median      3Q     Max 
-8.1917 -1.0729  0.8042  1.9021  3.3083 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  20.1917     0.7592  26.596   <2e-16 ***
typeSelf     -2.6167     1.0737  -2.437   0.0214 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.94 on 28 degrees of freedom
Multiple R-squared:  0.175,	Adjusted R-squared:  0.1455 
F-statistic:  5.94 on 1 and 28 DF,  p-value: 0.02141
```

Bringing it all together!

Self pollinated maize plants were on average 17.6 inches high, while the cross-pollinated plants had a height of 20.2 inches – a difference of 2.6 inches which was statistically significant ($F_{1,28}$ = 5.9, P = 0.02, R^2 = 0.15).

This version of our write-up has:

- Sample size information (degrees of freedom)
- Test statistic (F)
- Together these produce our *P* value
- Information on the variance explained by our model R^2
- Accurate information on our observations

```
Call:
lm(formula = height ~ type, data = darwin)

Residuals:
    Min      1Q  Median      3Q     Max 
-8.1917 -1.0729  0.8042  1.9021  3.3083 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  20.1917     0.7592  26.596   <2e-16 ***
typeSelf     -2.6167     1.0737  -2.437   0.0214 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.94 on 28 degrees of freedom
Multiple R-squared:  0.175,	Adjusted R-squared:  0.1455 
F-statistic:  5.94 on 1 and 28 DF,  p-value: 0.02141
```

Next Time

- We will look at more assumptions of our linear models and how to test we have a *good fit*
- Confidence intervals how to capture the *importance* or *effect size* of our models

Remember we have a discussion boards for:

- R code
- GitHub
- Stats theory

This week's assignments

1) Join the GitHub Classroom – Task 3 Wk 1

2) Complete last week's (Week 2) workshop:

Philip-Leftwich/**5023Y-Week2-Statistics**

3) Start this week's assignment

Philip-Leftwich/**5023Y-Week3-Statistics**