ANOVA Tables

What is an ANOVA Table?

Background

- In this presentation, we dive into ANOVA tables, a powerful tool for analysing differences.

What's an ANOVA table?

We use an Analysis of Variance (ANOVA) table as a descriptive tool to see how much variability in the outcome that different variables account for.

In an ANOVA, the total variability in the data is split into two components:

- The variation due to the differences between group means (**accounted-for variation**)
- The variation within the groups themselves (**a mix of the unaccounted-for variation and the natural variability**).

This division allows for assessing whether the means of different groups are substantially different from each other while accounting for the variability within each group,

The Data: In Brief

- The data we use to illustrate ANOVA tables gives the yield in metric tons per hectare of 58 varieties of soybeans.
- These soybeans are grown in four locations across two consecutive years in Australia.
- A data frame is given in R (under `agridat::australia.soybean`) with 464 observations.
- The relevant variables we look at in this example are:
  - **yield** - yield, metric tons / hectare
  - **env** - environment: 8 levels (a combination of the four locations in the two years)
  - **gen** - genotype of soybeans: 58 levels for 58 soybeans.
  - **oil** - oil (percentage)

What's an ANOVA table?

- ANOVA tables are used to investigate variation. It shows how much variability is accounted for by a variable.
- We first input the numerical variable, oil:

	Df	Sum Sq	Mean Sq	F value
oil	1	75.13	75.13	185.74
Residuals	462	186.87	0.40	
Total	463	262.00		

What's an ANOVA table?

- ANOVA tables are used to investigate variation. It shows how much variability is accounted for by a variable.
- First input the numerical variable, oil:

Effect of oil

	Df	Sum Sq	Mean Sq	F value
oil	1	75.13	75.13	185.74
Residuals	462	186.87	0.40	
Total	463	262.00		

Leftover

Degrees of Freedom

Degrees of freedom are the “number of choices” we have

Numerical Variable = 1

	Df	Sum Sq	Mean Sq	F value
oil	1	75.13	75.13	185.74
Residuals	462	186.87	0.40	
Total	463	262.00		

Leftover (still)

Total rows - 1

Sums of Squares

- Measures how much a line of best fit in oil varies from the overall average

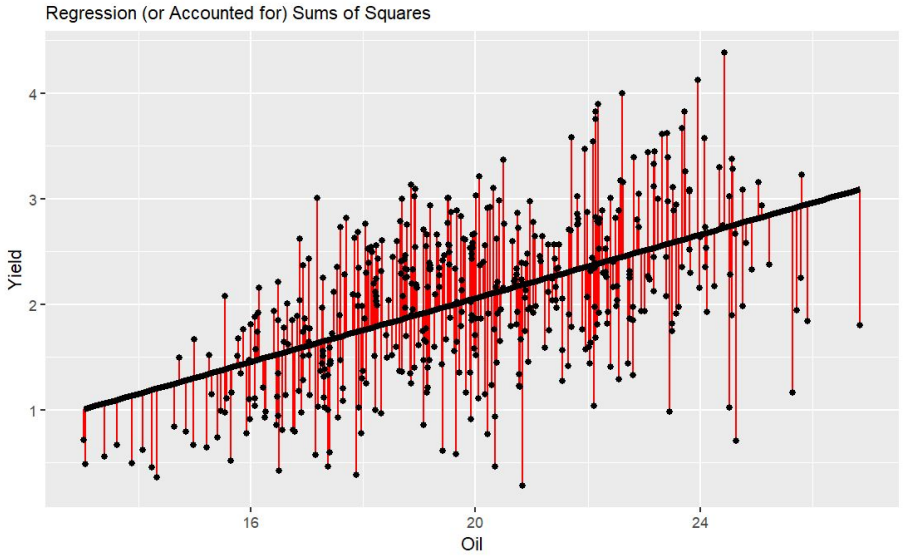
Sums of Squares

	Df	Sum Sq	Mean Sq	F value
oil	1	75.13	75.13	185.74
Residuals	462	186.87	0.40	
Total	463	262.00		

Sums of Squares

Visual representation of SS: Oil Sums of Squares

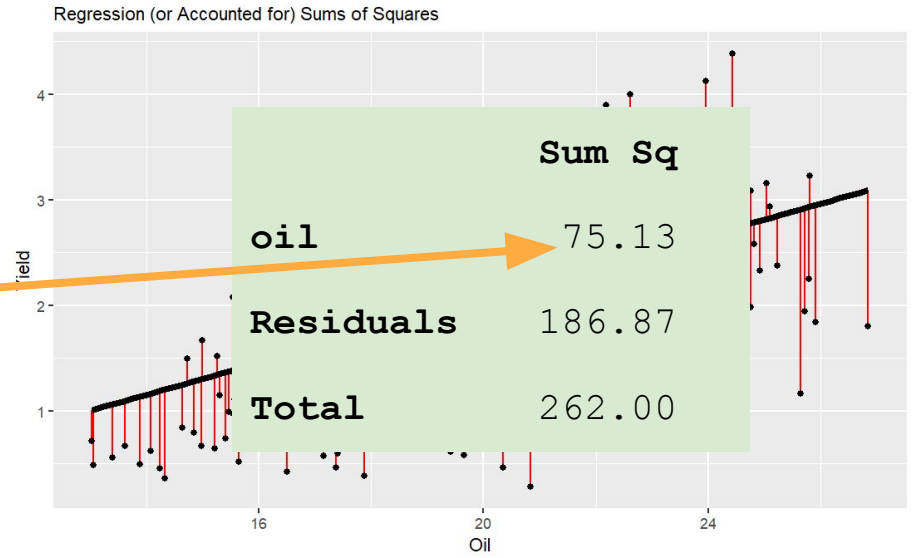
1. Plot Oil against Yield
2. Fit a line of best fit
3. This looks at the distance (squared) from Oil to the line of best fit.
4. In our case, this value is `75.13`



Sums of Squares

Visual representation of SS: Oil Sums of Squares

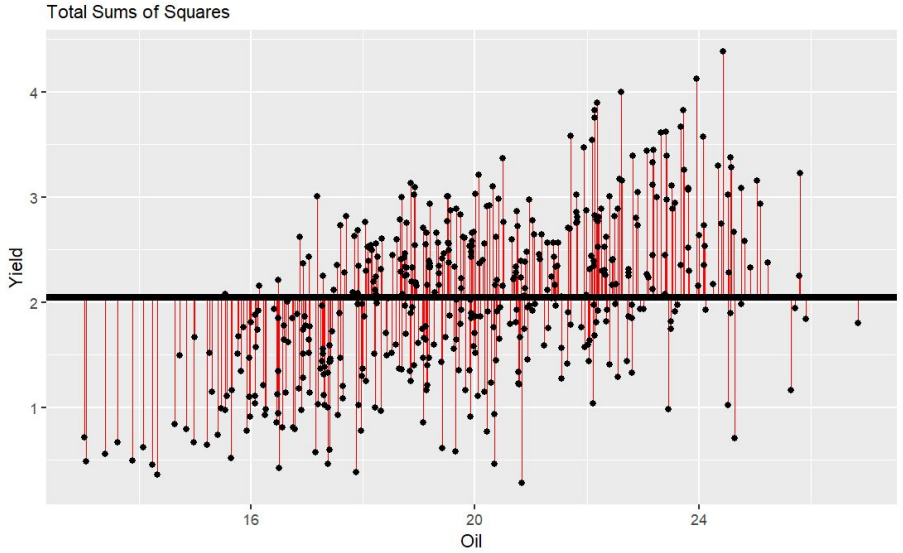
1. Plot Oil against Yield
2. Fit a line of best fit
3. This looks at the distance (squared) from Oil to the line of best fit.
4. In our case, this value is `75.13`



Sums of Squares

Visual representation of SS: Total Sums of Squares

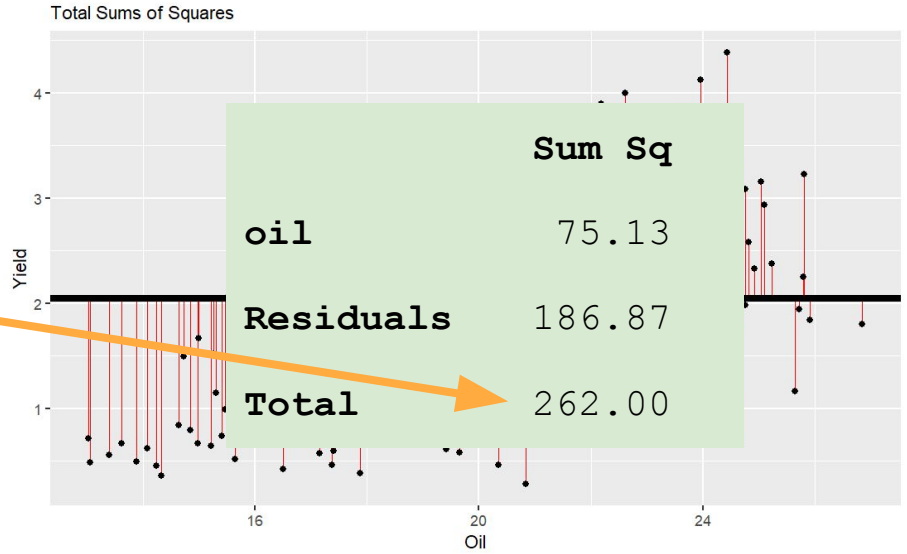
1. Plot Oil against Yield
2. Fit a line at the mean (black line)
3. This looks at the distance (squared) from Oil to the mean.
4. In our case, this value is `262.00`



Sums of Squares

Visual representation of SS: Total Sums of Squares

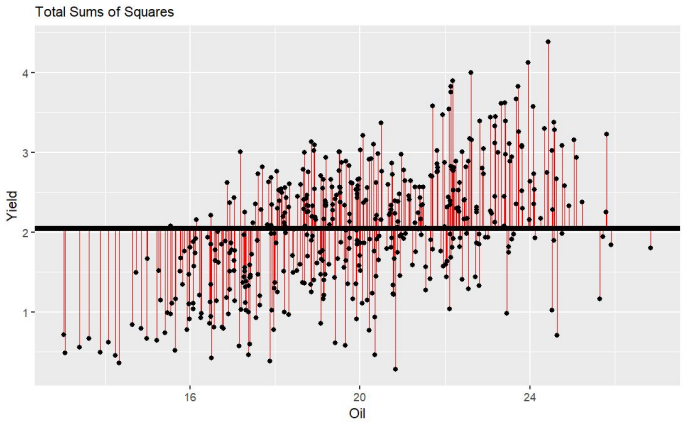
1. Plot Oil against Yield
2. Fit a line at the mean (black line)
3. This looks at the distance (squared) from Oil to the mean.
4. In our case, this value is `262.00`



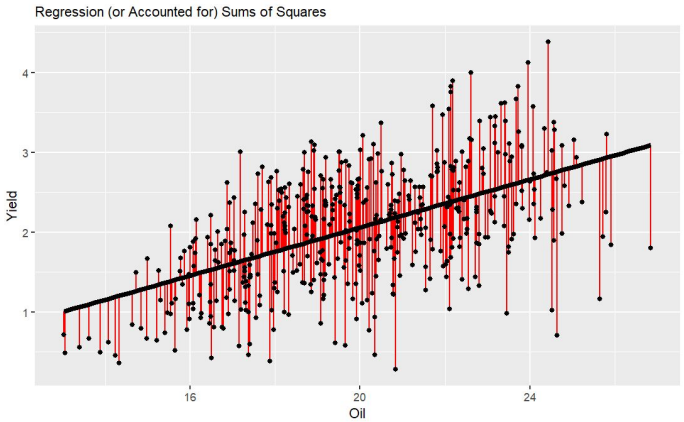
Sums of Squares

Visual representation of SS: Residual Sums of Squares

This is the “leftover” bit.



=

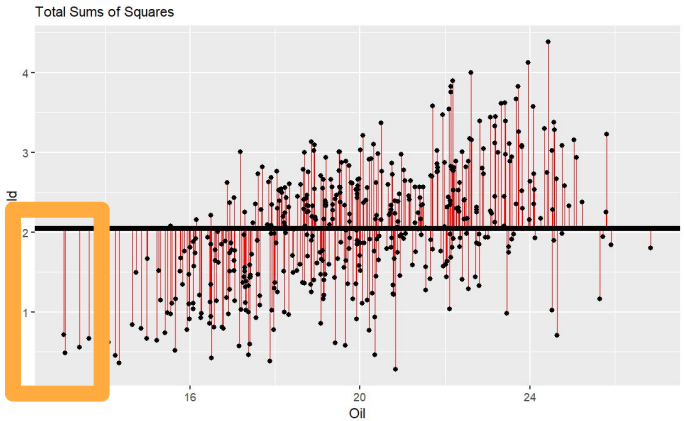


+ ?

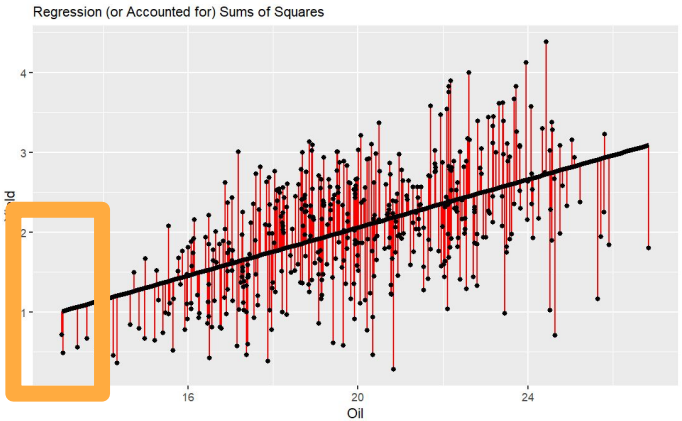
Sums of Squares

Visual representation of SS: Residual Sums of Squares

Let's focus on this bit



=

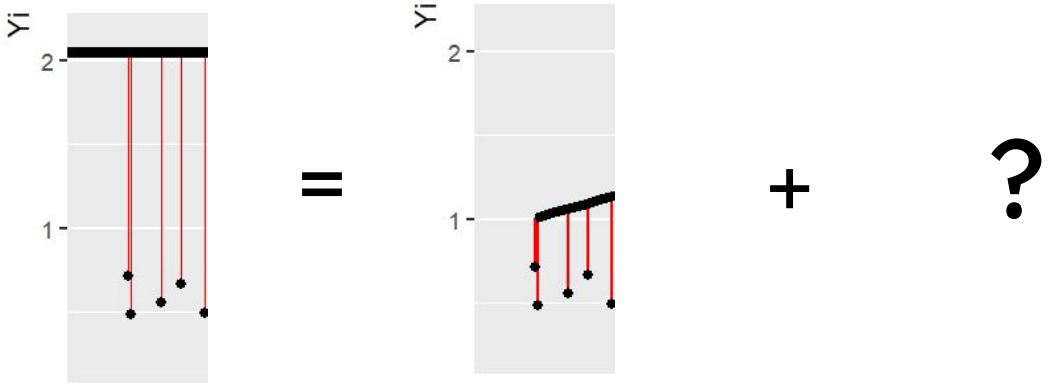


+ ?

Sums of Squares

Visual representation of SS: Residual Sums of Squares

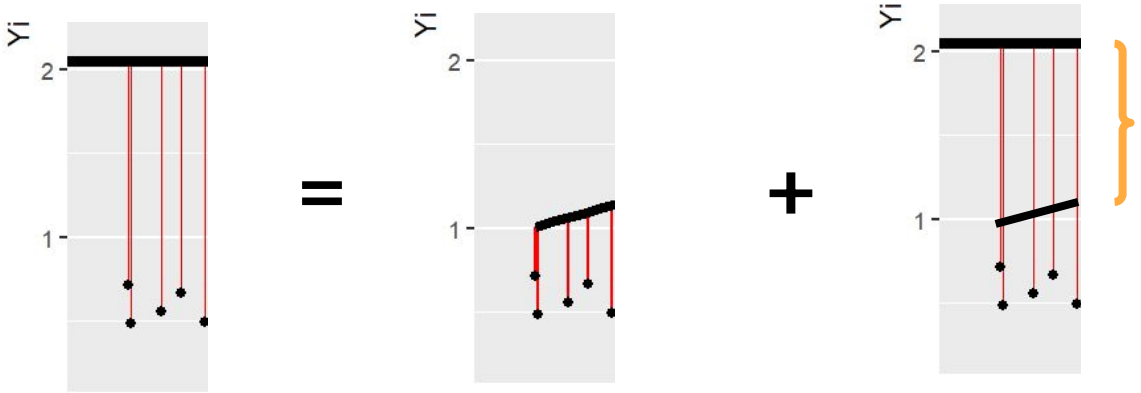
Let’s focus on this bit



Sums of Squares

Visual representation of SS: Residual Sums of Squares

Let's focus on this bit



The difference between the line of best fit and the mean

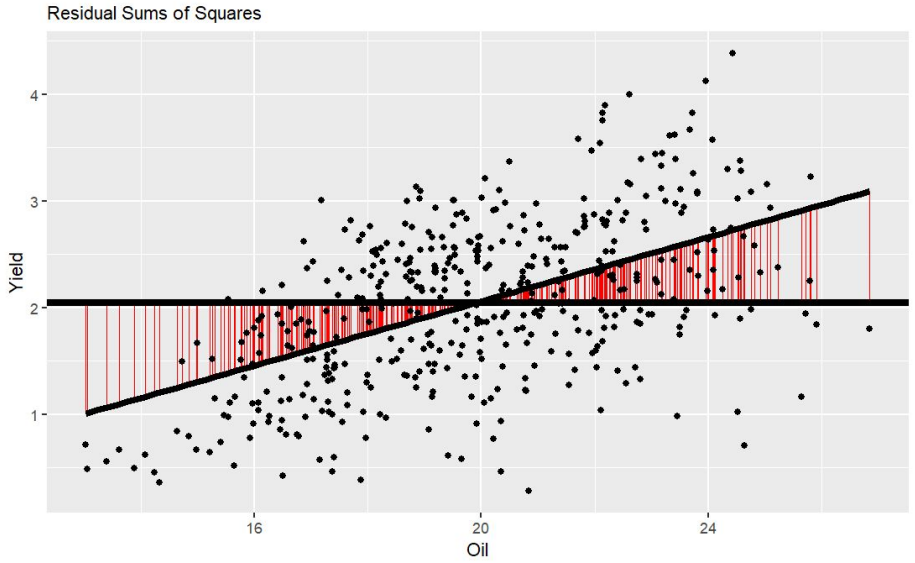
We haven't accounted for this variability yet.

Sums of Squares

Visual representation of SS: Residual Sums of Squares

1. The squared-difference between the line of best fit and the mean
2. In our case, this value is `186.87`

(Total SS - Accounted-for SS)

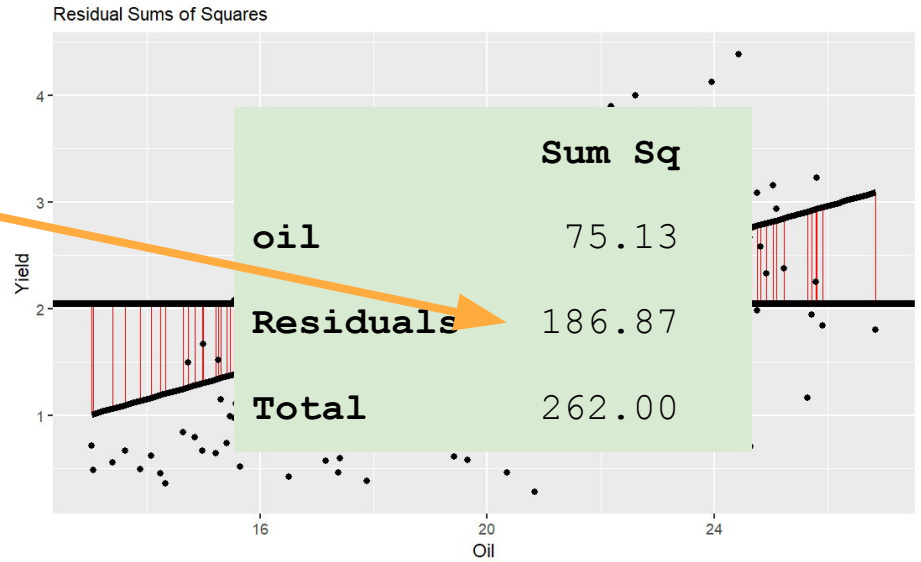


Sums of Squares

Visual representation of SS: Residual Sums of Squares

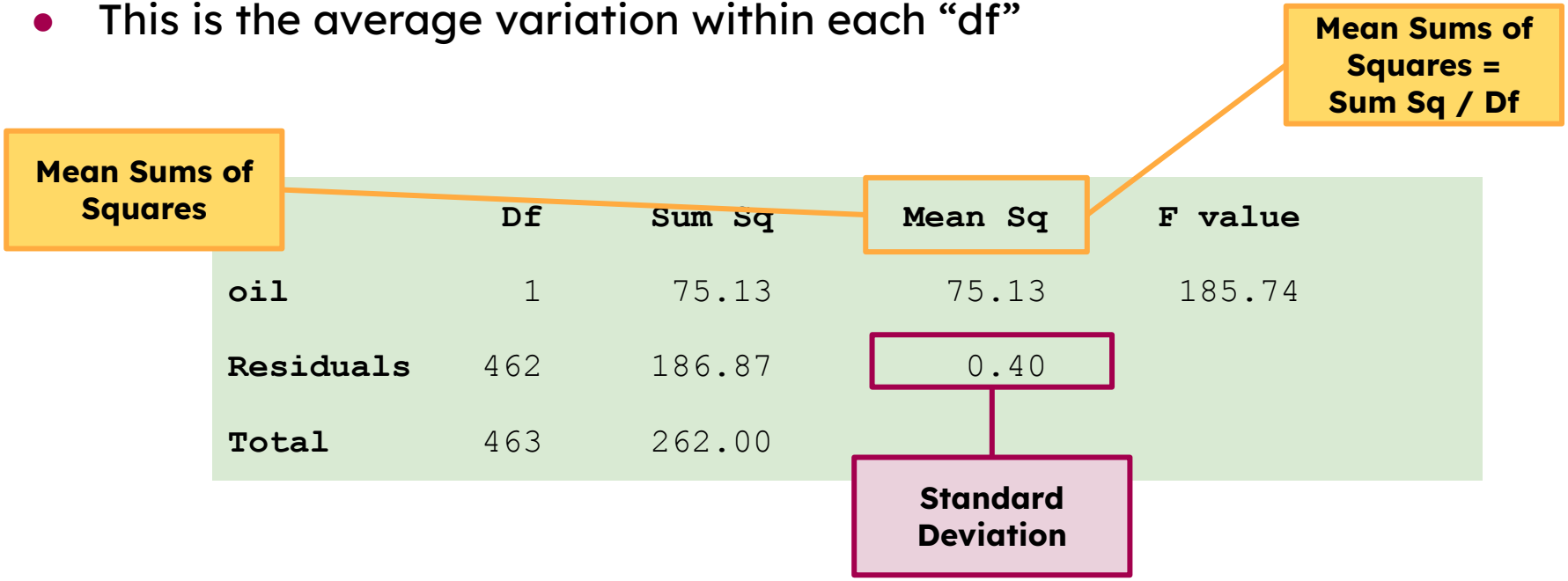
1. The squared-difference between the line of best fit and the mean
2. In our case, this value is `186.87`

(Total SS - Accounted-for SS)



Mean Sums of Squares

- This is the average variation within each “df”



	Df	Sum Sq	Mean Sq	F value
oil	1	75.13	75.13	185.74
Residuals	462	186.87	0.40	
Total	463	262.00		

F-Value

- This is the the variance between the “group” means by the variance in the residuals
- This helps ascertain whether the observed variability in the outcome (yield) are due to random chance or due to oil itself.

	Df	Sum Sq	Mean Sq	F value
oil	1	75.13	75.13	185.74
Residuals	462	186.87	0.40	
Total	463	262.00		

Mean Sq (oil) / Mean Sq (residuals

F-Value

- This is the the variance between the “group” means by the variance in the residuals
- This helps ascertain whether the observed variability in the outcome (yield) are due to random chance or due to oil itself.

	Df	Sum Sq	Mean Sq	F value
oil	1			185.74
Residuals	462			
Total	463			

This tells us: How much variation is accounted for by `oil` compared to the residuals.

Mean Sq (oil) / Mean Sq (residuals

F-Value

- This is the the variance between the “group” means by the variance in the residuals
- This helps ascertain whether the observed variability in the outcome (yield) are due to random chance or due to oil itself.

	Df	Sum Sq	Mean Sq	F value
oil	1			185.74
Residuals	462			
Total	463			

This is much greater than 1. What does that tell us? That a high amount of variability in the outcome is accounted for by `oil` compared to the residuals.

Mean Sq (oil) / Mean Sq (residuals

F-Value

- This is the the variance between the “group” means by the variance in the residuals
- This helps ascertain whether the observed variability in the outcome (yield) are due to random chance or due to oil itself.

	Df	Sum Sq	Mean Sq	F value
oil	1			185.74
Residuals	462			
Total	463			

What if this was approximately 1? Or less than 1?

Have a think. We look at this in Topic ten.

Mean Sq (oil) / Mean Sq (residuals

Interpretation

How much variability is accounted for by `oil`?

	Df	Sum Sq	Mean Sq	F value
oil	1	75.13	75.13	185.74
Residuals	462	186.87	0.40	
Total	463	262.00		

Interpretation

How much variability is accounted for by `oil`?

	Df	Sum Sq	Mean Sq	F value
oil	1	75.13	75.13	185.74
Residuals	462	186.87	0.40	
Total	463	262.00		

`75.13/262.00 = 0.2867557...`

28.6% of variability in yield is accounted for by oil.

Interpretation

How much variability is accounted for by `oil`?

	Df	Sum Sq	Mean Sq	F value
oil	1	75.13	75.13	185.74
Residuals	462	186.87	0.40	
Total	463	262.00		

`75.13/262.00 = 0.2867557...`

28.6% of variability in yield is accounted for by oil.

71.4% of variability is unaccounted for. But, we can add more variables to see...

ANOVA and factors

We can look at another variable. Here, have a factor

	Df	Sum Sq	Mean Sq	F value
loc	3	17.32	5.77	15.63
Residuals	459	169.55	0.37	
Total	463	262.00		

ANOVA and factors

We can look at another variable. Here, have a factor

	Df	Sum Sq	Mean Sq	F value
loc	3	25.61	8.54	16.61
Residuals	460	236.40	0.51	
Total	463	262.00		

Effect of environment

Leftover (still)

ANOVA and factors

We can look at another variable. Here, have a factor

Number of factor levels - 1

	Df	Sum Sq	Mean Sq	F value
loc	3	25.61	8.54	16.61
Residuals	460	236.40	0.51	
Total	463	262.00		

Leftover (still)

ANOVA and factors

We can look at another variable. Here, have a factor

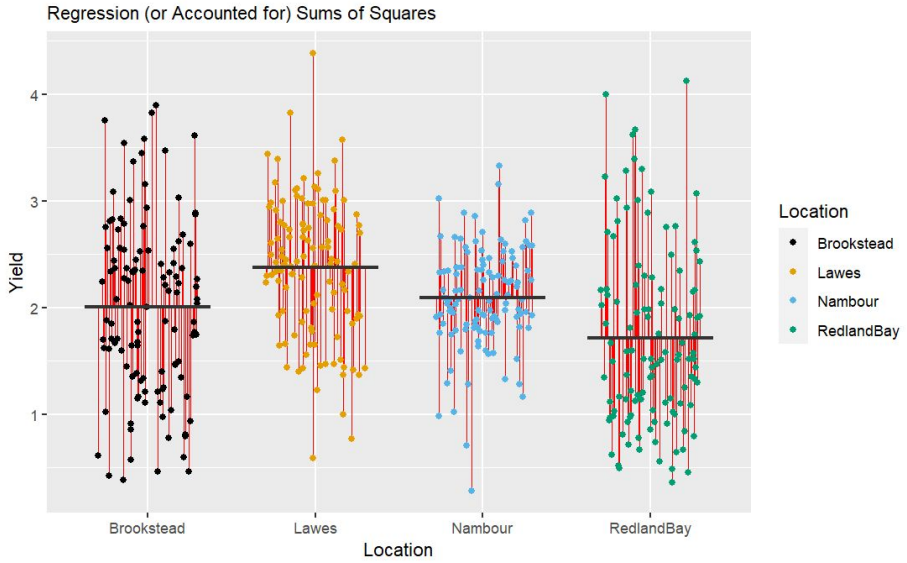
	Df	Sum Sq	Mean Sq	F value
loc	3	25.61	8.54	16.61
Residuals	460	236.40	0.51	
Total	463	262.00		

Sums of Squares

Sums of Squares (factor)

Visual representation of SS: Loc Sums of Squares

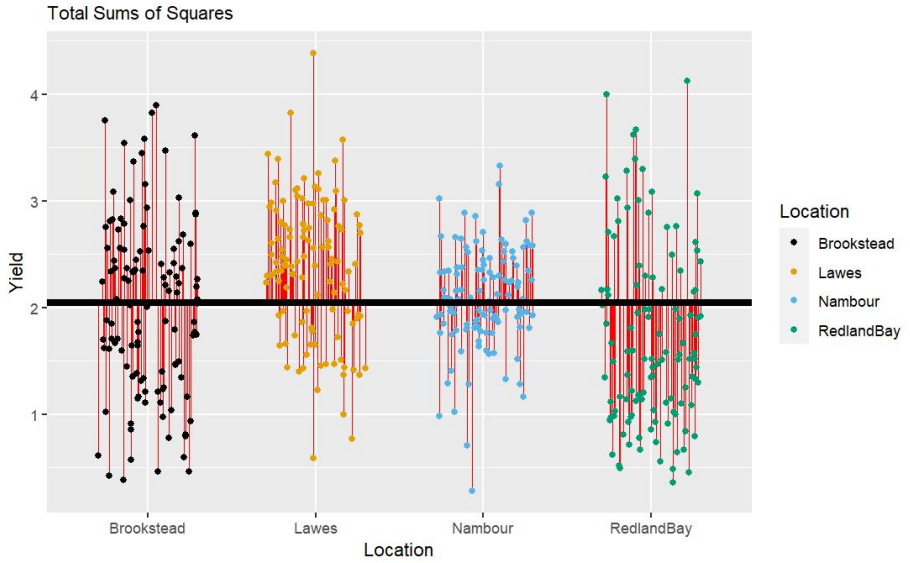
1. Plot location (loc) against Yield
2. Find the mean Yield for each “Loc” group
3. This looks at the distance (squared) from the yield to the mean yield for that loc group



Sums of Squares (factor)

Visual representation of SS: Total Sums of Squares

1. Plot Loc against Yield
2. Fit a line at the mean (black line)
3. This looks at the distance (squared) from loc to the mean.



Sums of Squares (factor)

Visual representation of SS: Residual Sums of Squares

The squared-difference between the group means and the overall mean

(Total SS - Accounted-for SS)

ANOVA and factors

We can look at another variable. Here, have a factor

Calculated the same as before

	Df	Sum Sq	Mean Sq	F value
loc	3	25.61	8.54	16.61
Residuals	460	236.40	0.51	
Total	463	262.00		

ANOVA and factors

We can look at another variable. Here, have a factor

Calculated the same as before

	Df	Sum Sq	Mean Sq	F value
loc	3	25.61	8.54	16.61
Residuals	460	236.40	0.51	
Total	463	262.00		

`25.61/262.00 = 0.09774809...`

9.77% of variability in yield is accounted for by location.

ANOVA and factors

How about having the two variables in there together:

	Df	Sum Sq	Mean Sq	F value
oil	1	75.13	75.13	203.39
loc	3	17.32	5.77	15.63
Residuals	459	169.55	0.37	
Total	463	262.00		

Interpretation

How much variability is accounted for now?

	Df	Sum Sq	Mean Sq	F value
oil	1	75.13	75.13	203.39
loc	3	17.32	5.77	15.63
Residuals	459	169.55	0.37	
Total	463	262.00		

```
(75.13 + 17.32)/262.00 = 0.35286...
```

35.3% of variability in yield is accounted for by oil and location.

Interpretation

How much variability is accounted for now?

	Df	Sum Sq	Mean Sq	F value
oil	1	75.13	75.13	203.39
loc	3	17.32	5.77	15.63
Residuals	459	169.55	0.37	
Total	463	262.00		

`(75.13 + 17.32)/262.00 = 0.35286...`

35.3% of variability in yield is accounted for by oil and location.

64.7% of variability in yield is still unaccounted for. We can keep adding variables!

Unidentified variability

- 35.3% of variability in yield is accounted for by oil and location.
- The other 64.7% is unaccounted-for variability.
  This is split into:
  - **Random (natural) variability.**
  - **Unidentified variability.**
    This is variability that could be accounted for, but has not been, either because we do not have those variables in the data set, or because we have not put them in the data (other variables in the data set or interactions).

The aim is to reduce the unidentified variability.

Conclusion

- ANOVA tables help find real differences between groups.
- They're a key tool in exploring data to help determine variables that account for variability
- The aim is to reduce the unidentified variability.