More on categorical explanatory variables

Outline

Recap of last week

More than one categorical variable

Mixing categorical and continuous

Tips and tricks to reading outputs

Outline

Recap of last week

- EX1: How to choose a model

More than one categorical variable

- EX2: Two categorical variables
- EX3: Interactions

Mixing categorical and continuous

- EX4: Categorical and continuous

Tips and tricks to reading outputs

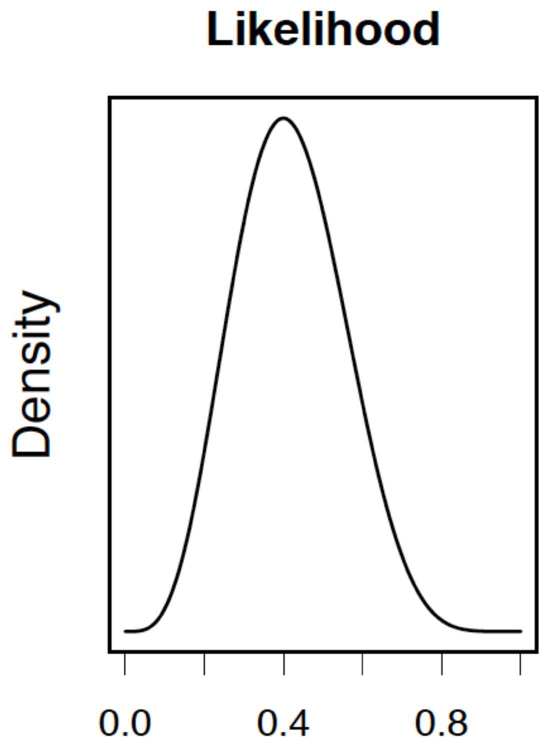
- EX5: What has been done?

Recap!

What have we covered so far?

What have we covered so far?

Began with Maximum Likelihood Estimation



What have we covered so far?

Began with Maximum Likelihood Estimation

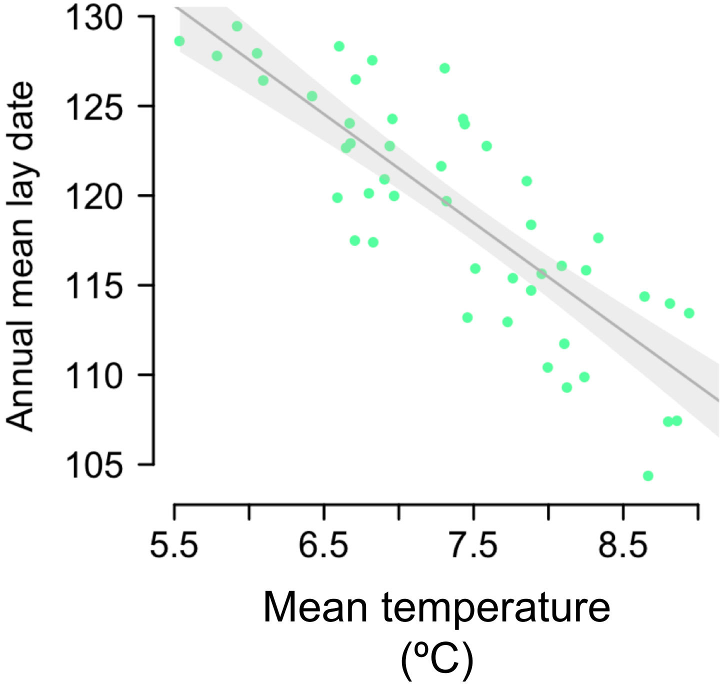
Then onto linear models $Y_i = \alpha + \beta X_i + \varepsilon_i$

What have we covered so far?

Began with Maximum Likelihood Estimation

Then onto linear models

Looked at continuous variables

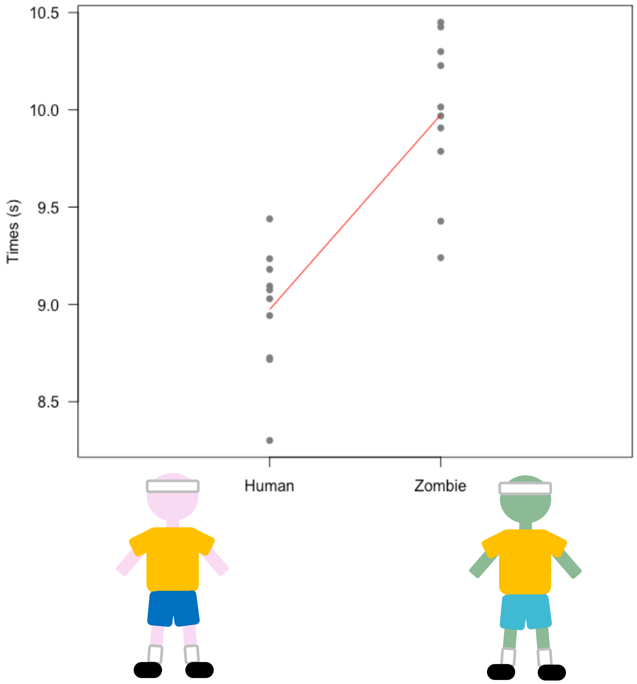


What have we covered so far?

Began with Maximum Likelihood Estimation

Then onto linear models

Looked at continuous variables and categorical



What have we covered so far?

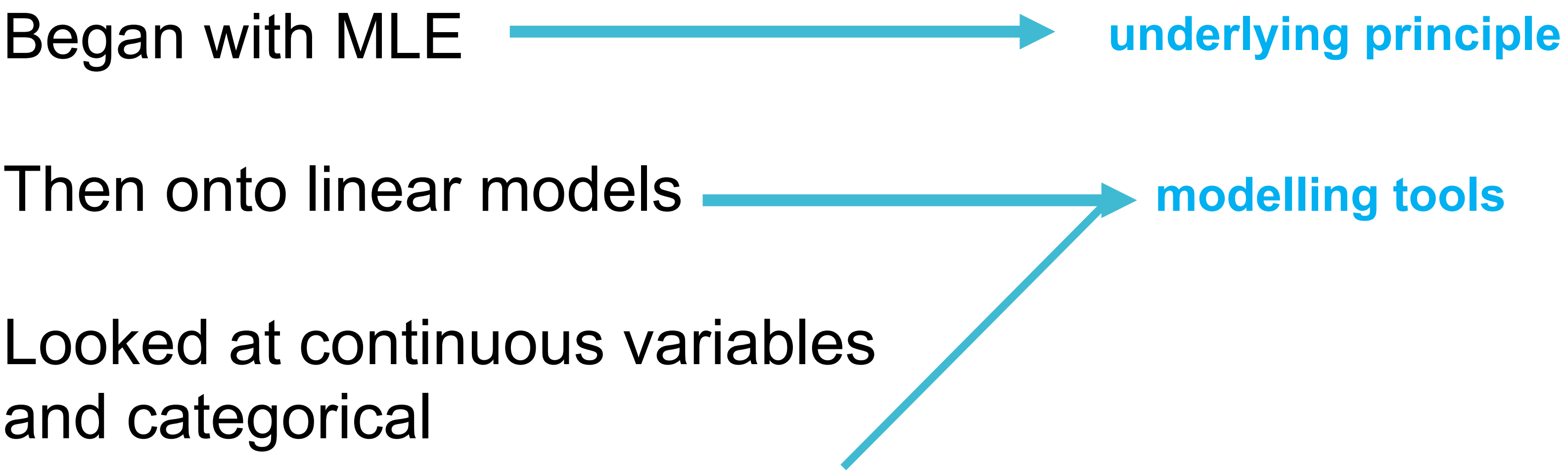
Began with MLE → **underlying principle**

Then onto linear models → **modelling tools**

Looked at continuous variables and categorical

Credit: commons.wikimedia

What have we covered so far?



Began with MLE → underlying principle

Then onto linear models → modelling tools

Looked at continuous variables and categorical → modelling tools

NEXT:

more tools…. This week = how to combine variables

Later = how to model when error is not normal

But why?

But why?

Aims of the course:

But why?

Aims of the course:

To be able to analyse own data

But why?

Aims of the course:

To be able to analyse own data

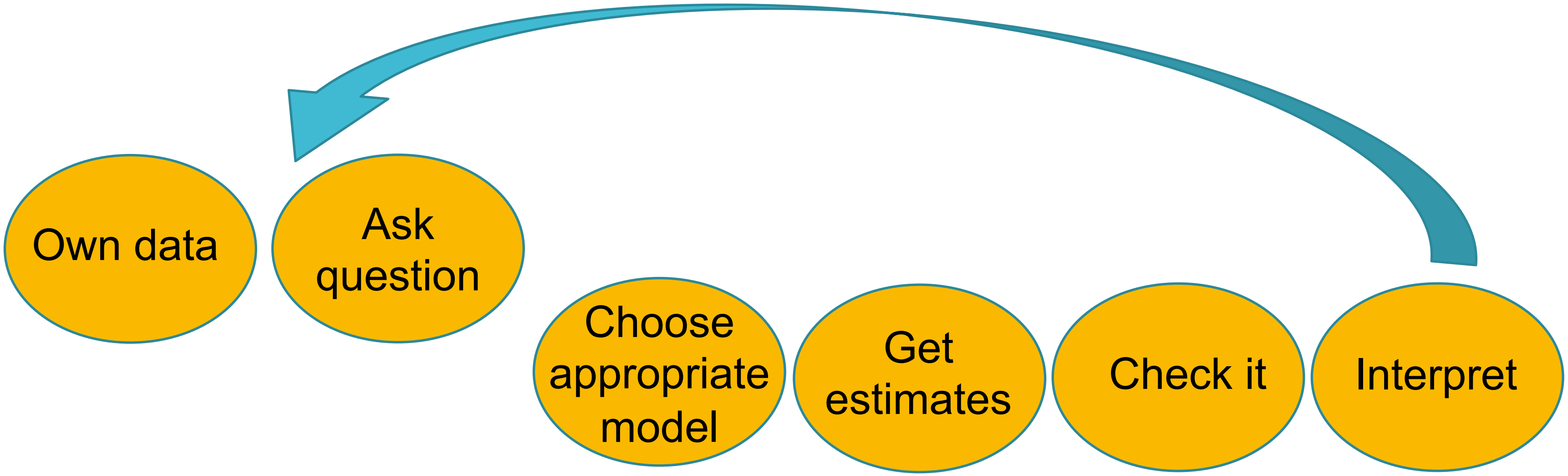
Giving you tools (some of that is programming, lots is the models)

But why?

Aims of the course:

To be able to analyse own data

Giving you tools (some of that is programming, lots is the models)

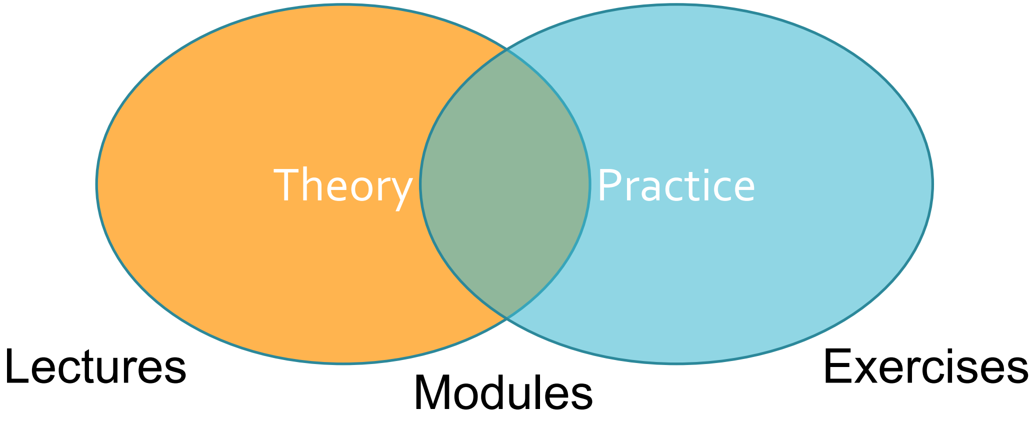


Links between theory and practice

Lectures tell you HOW tool works, and some of mathematical principles behind them

Exercises let you practice USING the tools

Two different sets of skills, but need both for statistics



Theory

Practice

Lectures

Modules

Exercises

Exercise 1: Choosing a model

- Complete Part A of the module

ANSWERS PART A

Dataset 1:

Dataset 2:

Dataset 3:

ANSWERS PART A

Dataset 1: categorical explanatory so…. differences in means

Dataset 2: continuous explanatory so .... relationship

Dataset 3: categorical explanatories so ….. differences in means and maybe interaction

ANSWERS PART A

Dataset 1: categorical explanatory so…. differences in means

Dataset 2: continuous explanatory so .... relationship

Dataset 3: categorical explanatories so ….. differences in means and maybe interaction

Maximum likelihood estimation of parameters

Last week

Looked at categorical explanatory variables

Using linear models

Finished with more than one variable

More than one categorical variable

Example from last week

Data on fertiliser treatments from Rothamsted

Four fertiliser treatments: *control, manure, fertilised, stopped*

Time: *before1970, after1970*

Example from last week

Could analyse both in separate models

```
lm(yield ~ Treatment, data = Rothamsted)
```

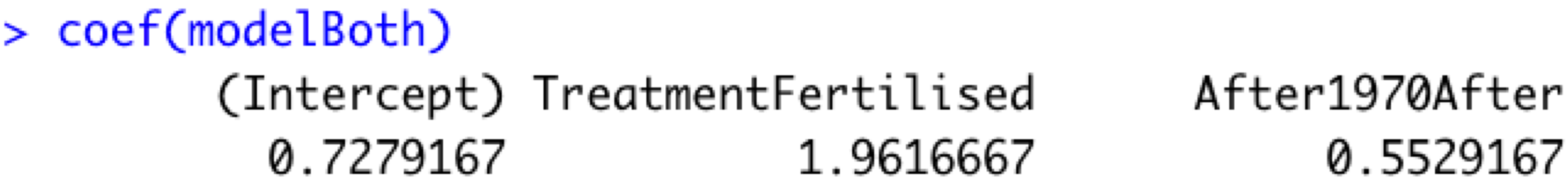
```
lm(yield ~ Time, data = Rothamsted)
```

Exercise 2: Two categorical explanatory variables

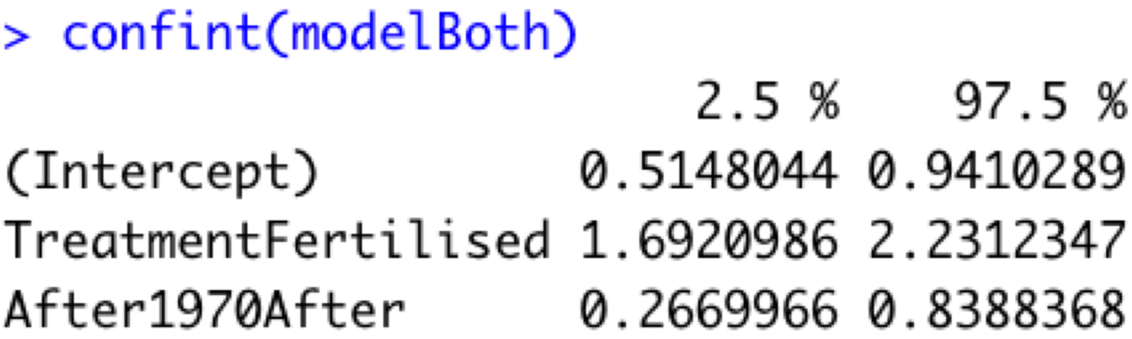
- Complete Part B of the module

Example from last week

Example from last week

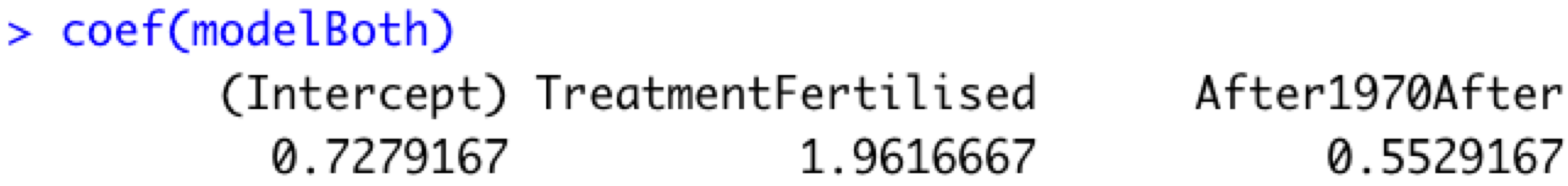


```
> coef(modelBoth)
        (Intercept) TreatmentFertilised       After1970After
          0.7279167           1.9616667            0.5529167
```

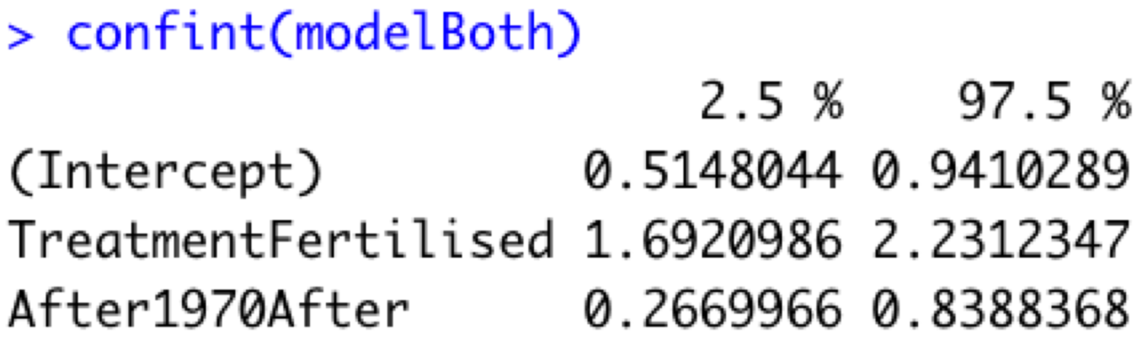


```
> confint(modelBoth)
                        2.5 %    97.5 %
(Intercept)         0.5148044 0.9410289
TreatmentFertilised 1.6920986 2.2312347
After1970After      0.2669966 0.8388368
```

Example from last week



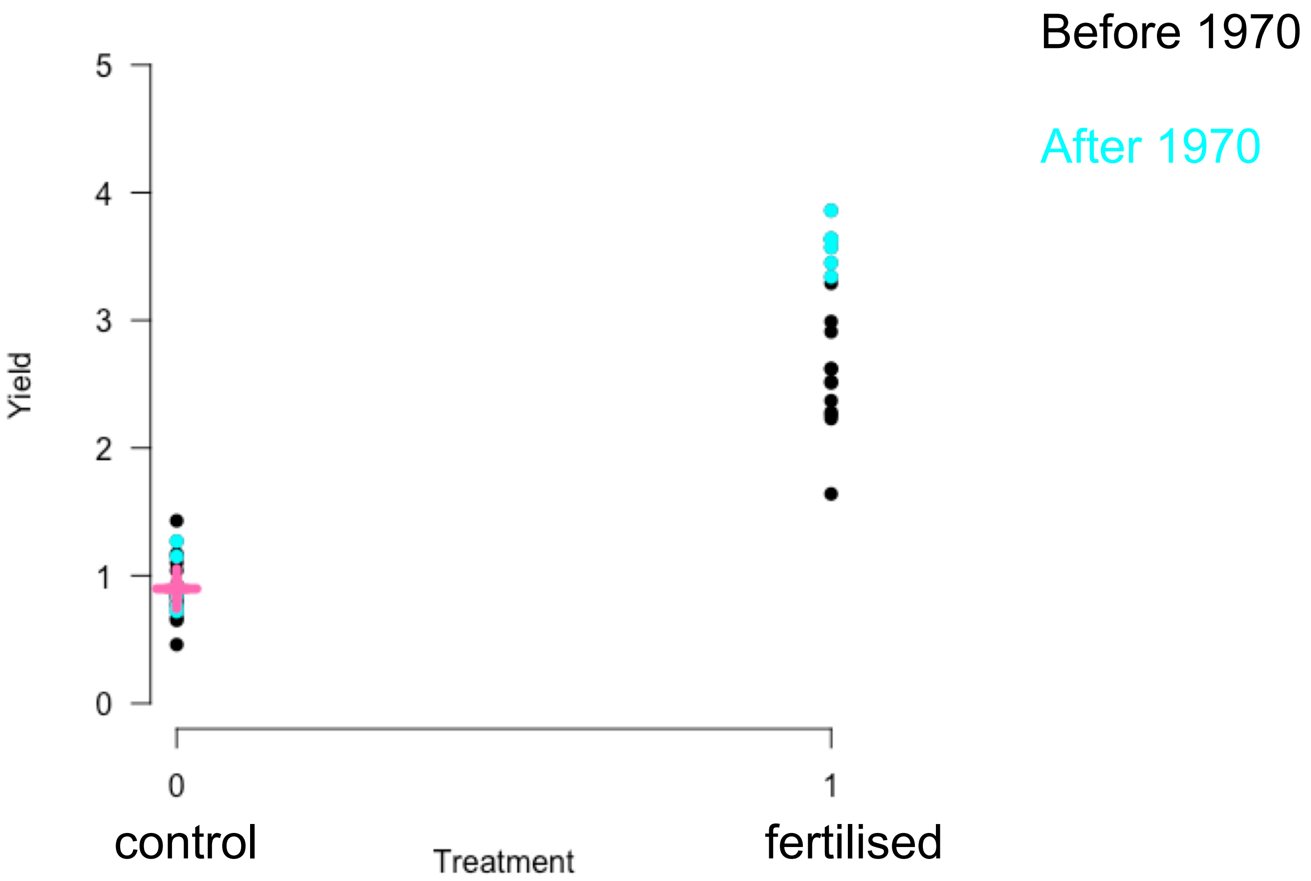
```
> coef(modelBoth)
        (Intercept) TreatmentFertilised       After1970After
          0.7279167           1.9616667            0.5529167
```



```
> confint(modelBoth)
                        2.5 %    97.5 %
(Intercept)         0.5148044 0.9410289
TreatmentFertilised 1.6920986 2.2312347
After1970After      0.2669966 0.8388368
```

$$Y_i = \alpha + \beta X_i$$

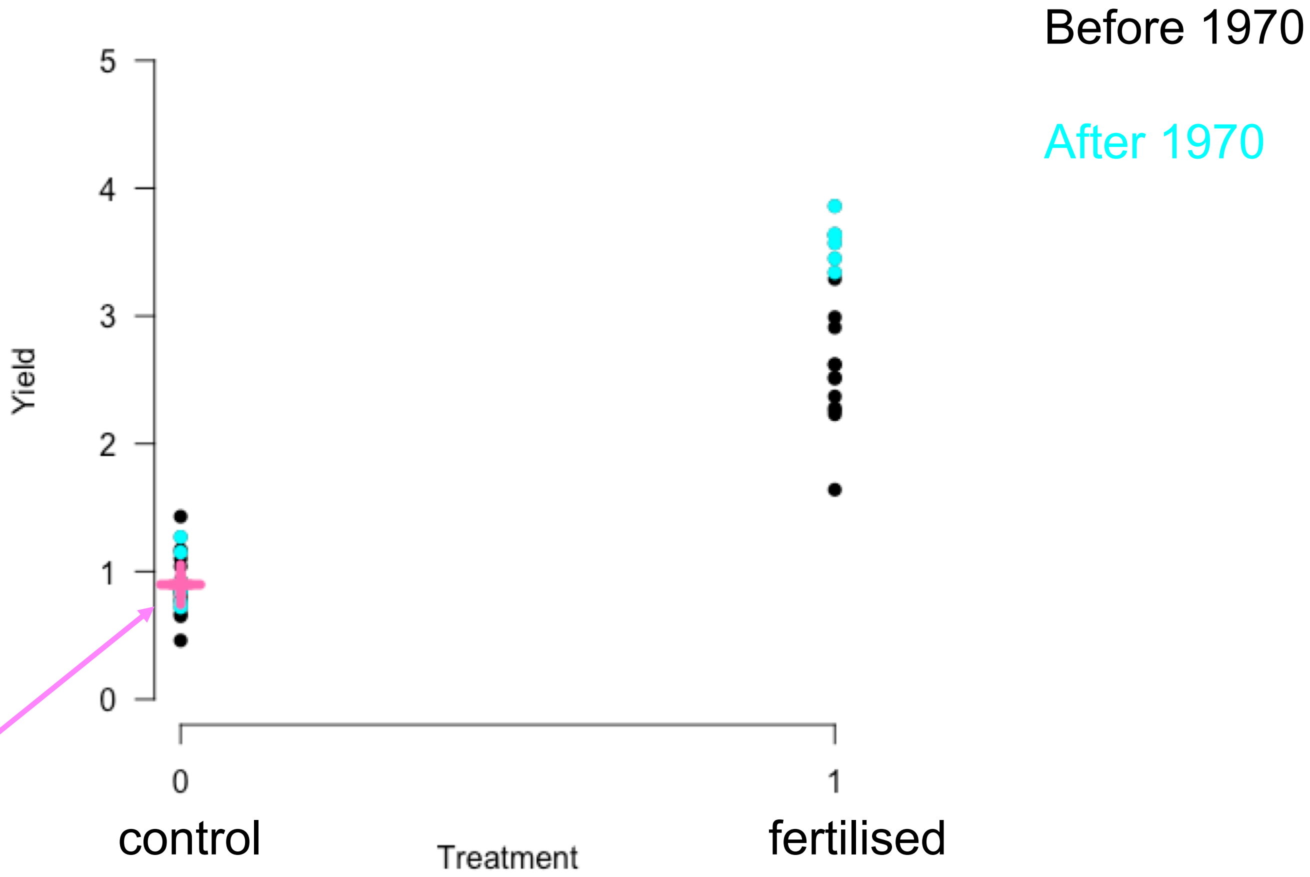
Example from last week



$$Y_i = \alpha + \beta X_i$$

```
> coef(modelBoth)
        (Intercept) TreatmentFertilised       After1970After
          0.7279167           1.9616667            0.5529167
```

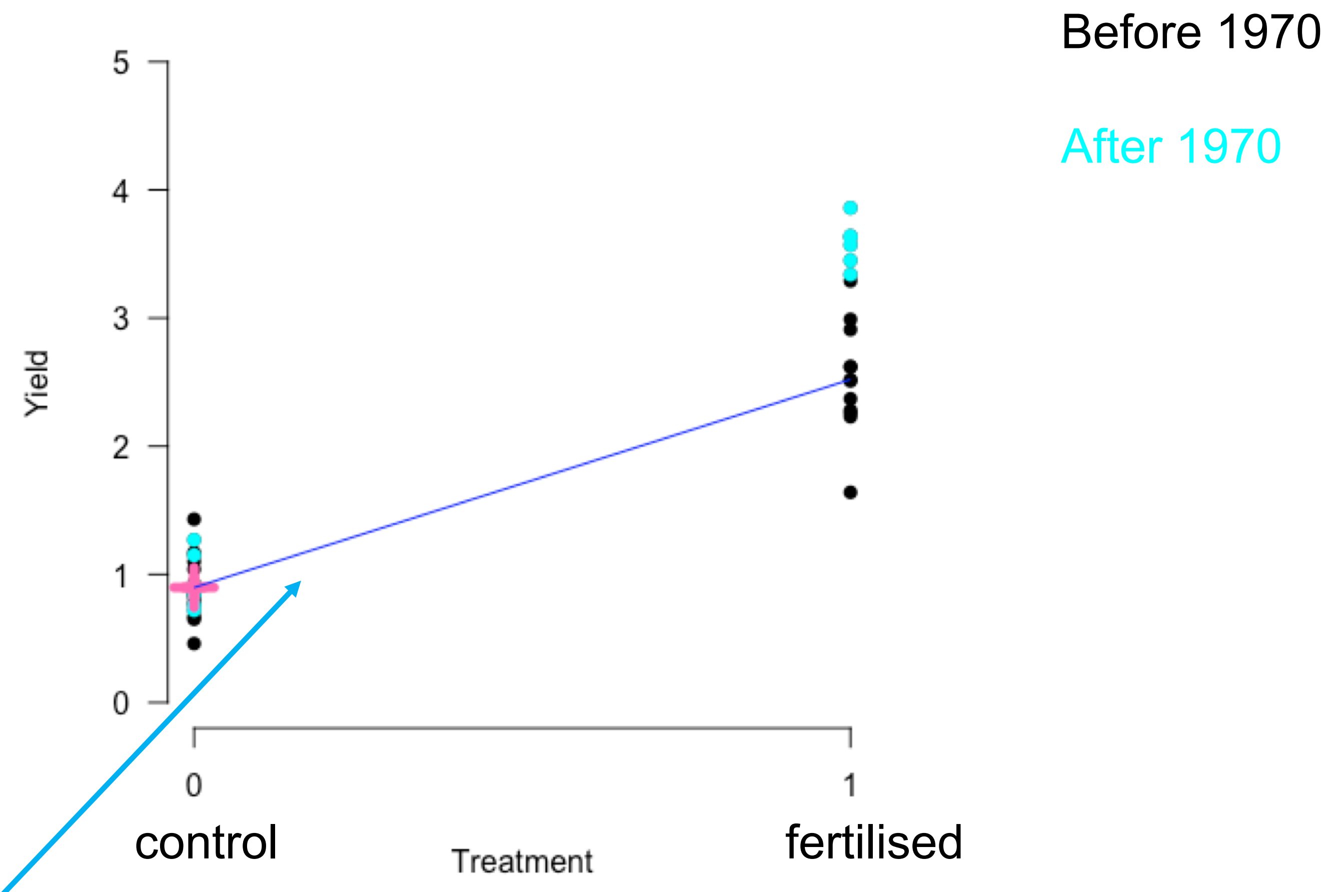
Example from last week



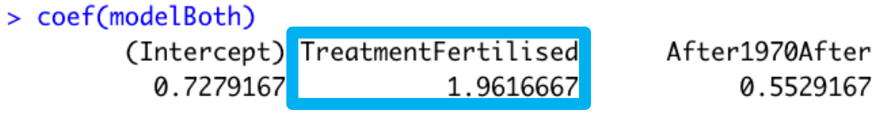
$$Y_i = \boxed{\alpha} + \beta X_i$$

```
> coef(modelBoth)
        (Intercept) TreatmentFertilised        After1970After 
          0.7279167           1.9616667             0.5529167 
```

Example from last week

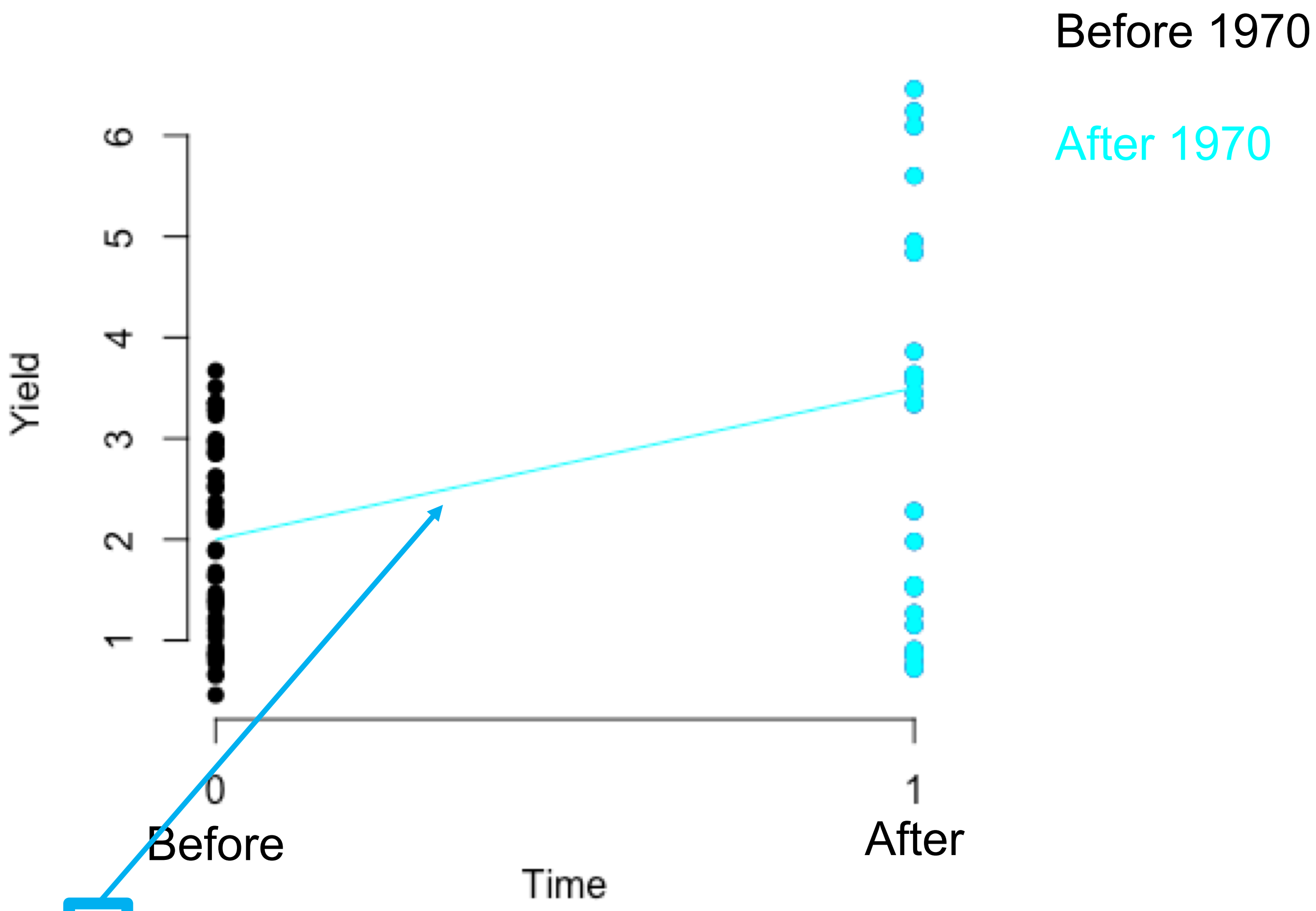


$$Y_i = \alpha + \boxed{\beta} X_i$$



```
> coef(modelBoth)
        (Intercept) TreatmentFertilised        After1970After
          0.7279167           1.9616667             0.5529167
```

Example from last week



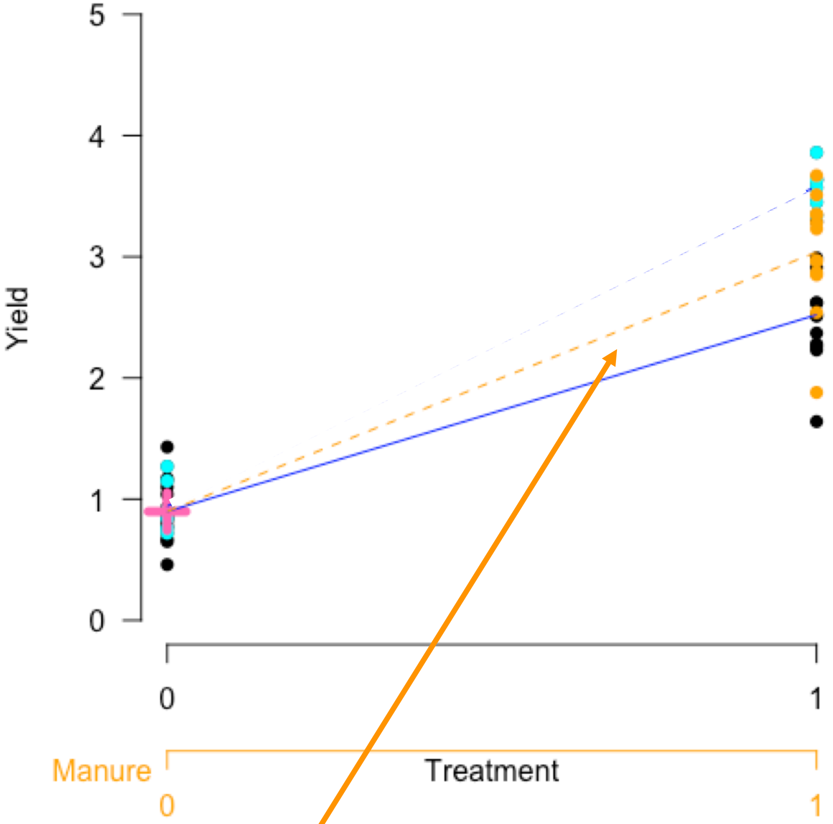
$$Y_i = \alpha + \beta X_i + \boxed{\beta}_2 X_{2_i}$$

```
> coef(modelBoth)
        (Intercept) TreatmentFertilised       After1970After
          0.7279167           1.9616667            0.5529167
```

One effect for all Treatments

What about more than one group?

What about more than one group?



Before 1970

After 1970

Manure before 1970

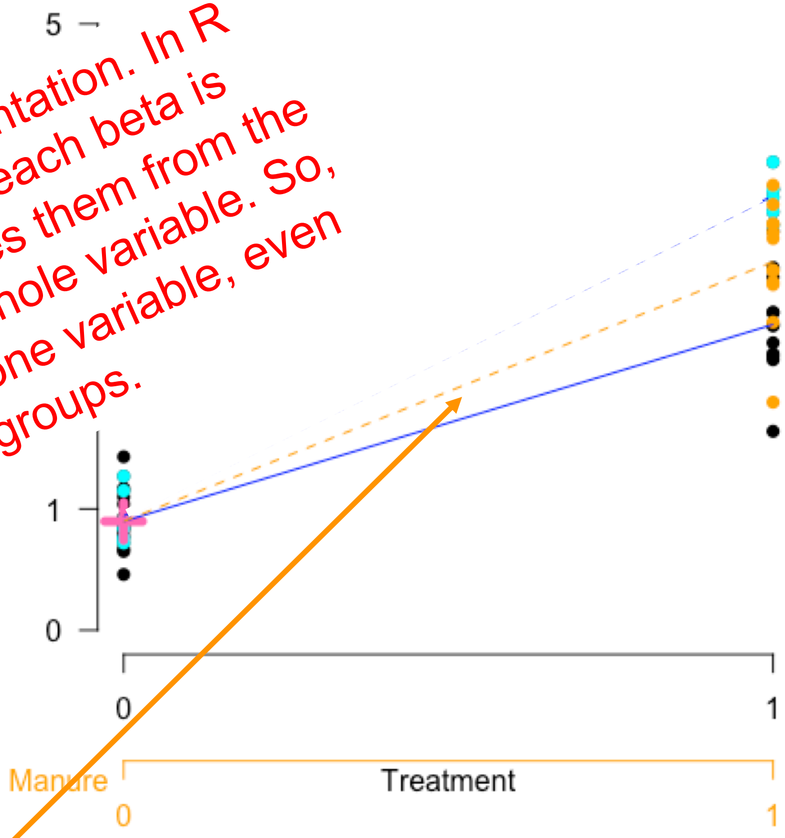
control

manure

$$Y_i = \alpha + \beta X_i + \beta_2 X_{2_i} + \boxed{\beta}_3 X_i$$

Different dimension

What about more than one group?



This is a visual representation. In R the standard error for each beta is the same. It calculates them from the variance from the whole variable. So, remember it is all one variable, even though there are groups.

Before 1970

After 1970

Manure before 1970

control

manure

$$Y_i = \alpha + \beta \boxed{X_i} + \beta_2 X_{2_i}$$

Summary

All about differences in means

Capture difference as a line with intercept and slope

Intercept = a group mean

Slope = difference between intercept group and others

Summary

So… we know what they values should mean

Did they add up?

Summary

So… we know what they values should mean

Did they add up? No

Summary

So… we know what they values should mean

Did they add up? No

Need interactions

Interactions

Why?

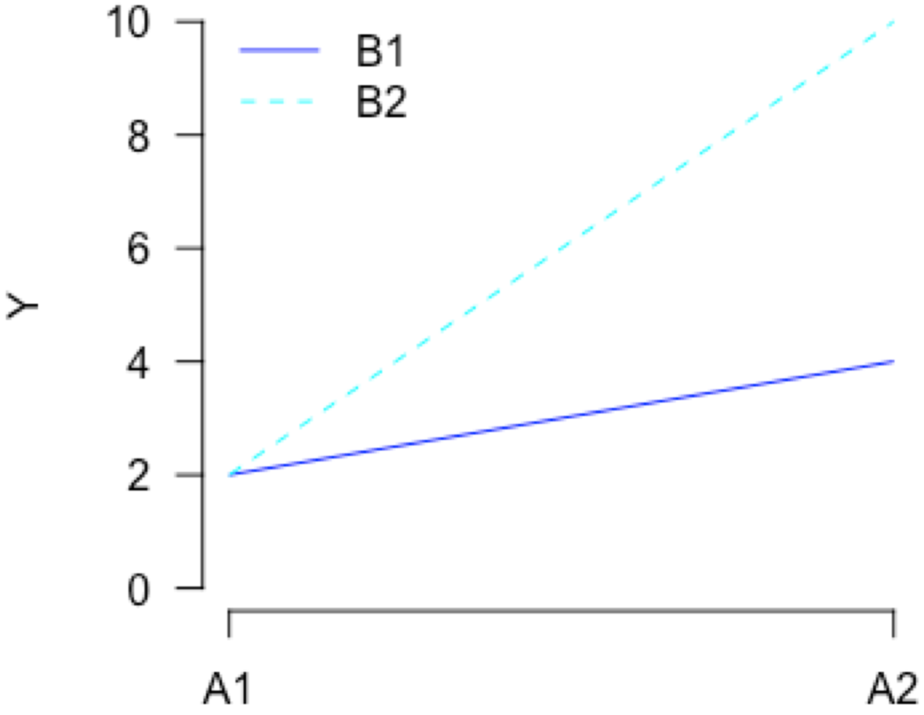
Why do we want to include them?

What do they tell us?

Why?

Why do we want to include them? Sometimes the effect of one variable depends on the effect of another

What do they tell us? How the effects change



How?

In the module

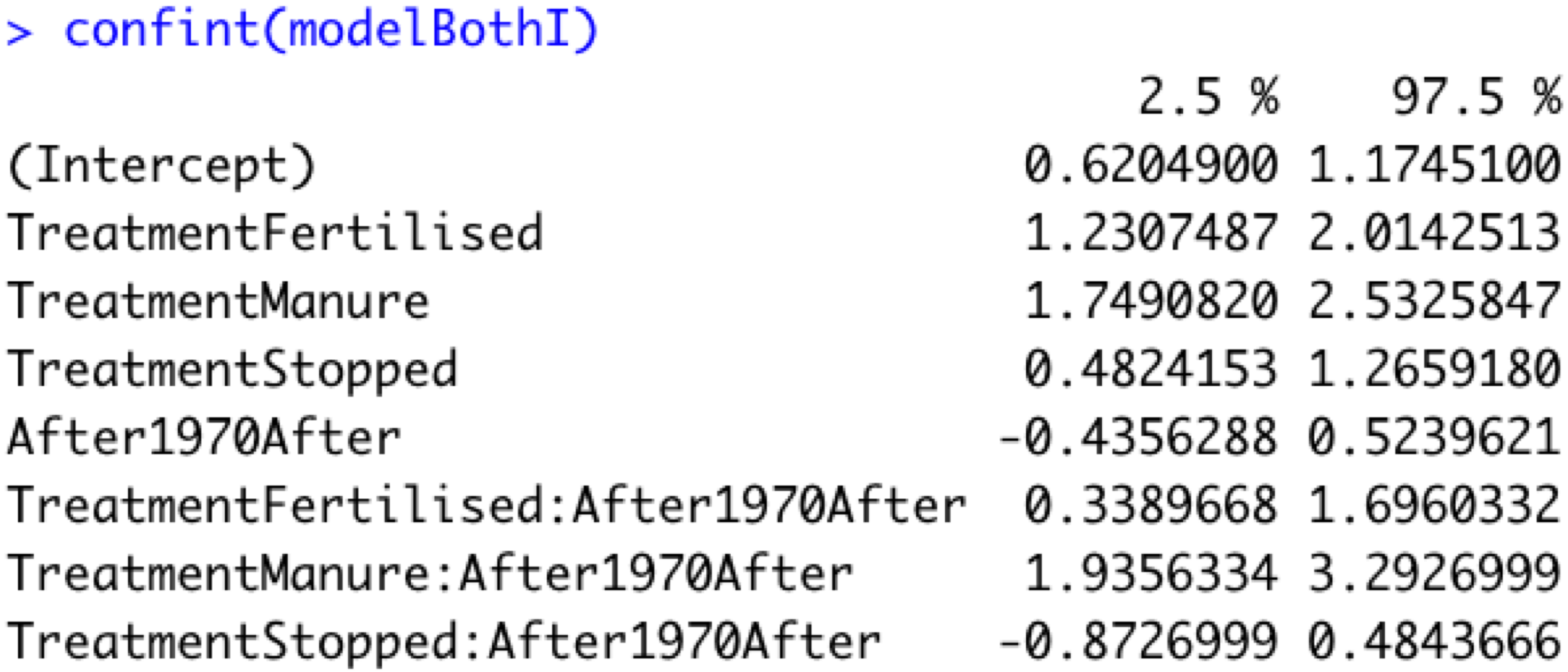
Want you to try interpretation on your own first

Exercise 3: Interactions

- Complete Part C of the module

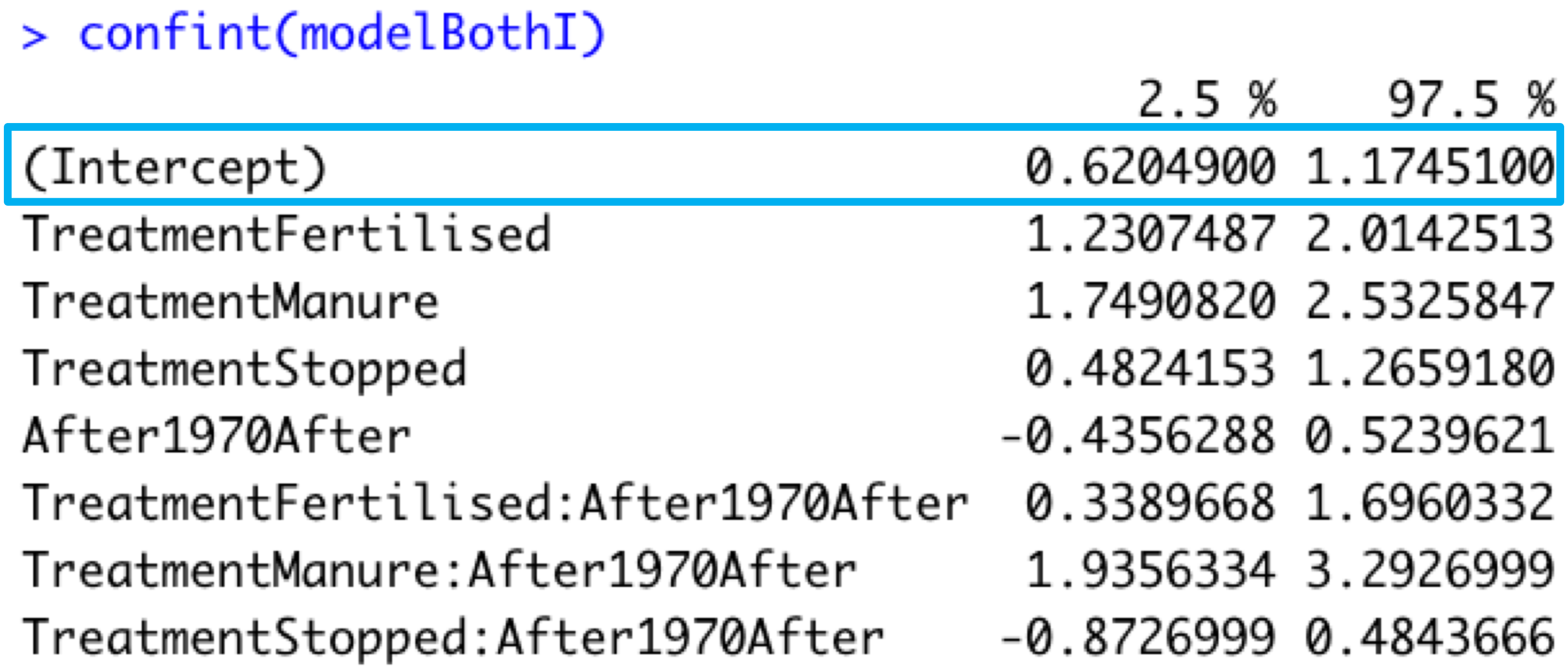
ANSWERS PART C

ANSWERS PART C



```
> confint(modelBothI)
                                      2.5 %    97.5 %
(Intercept)                       0.6204900 1.1745100
TreatmentFertilised               1.2307487 2.0142513
TreatmentManure                   1.7490820 2.5325847
TreatmentStopped                  0.4824153 1.2659180
After1970After                   -0.4356288 0.5239621
TreatmentFertilised:After1970After  0.3389668 1.6960332
TreatmentManure:After1970After    1.9356334 3.2926999
TreatmentStopped:After1970After  -0.8726999 0.4843666
```

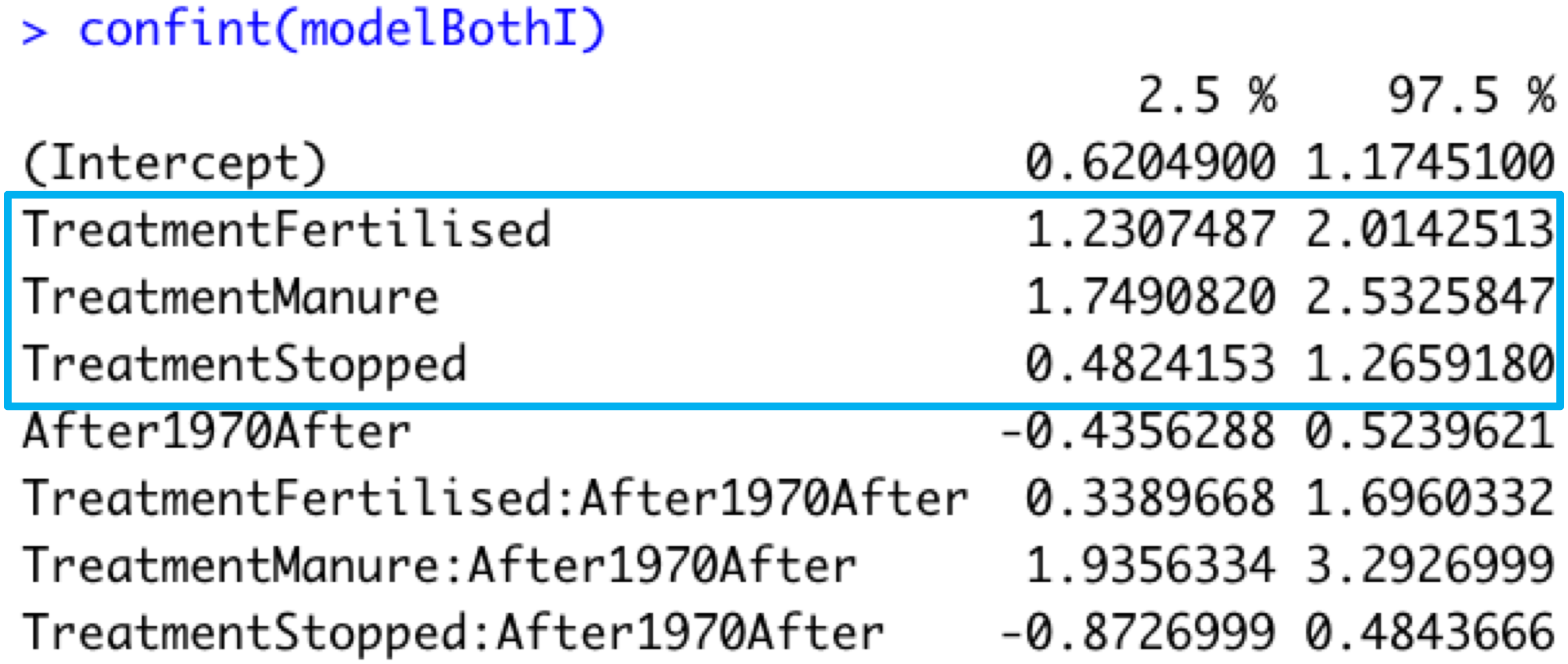
ANSWERS PART C



```
> confint(modelBothI)
                                     2.5 %    97.5 %
(Intercept)                      0.6204900 1.1745100
TreatmentFertilised              1.2307487 2.0142513
TreatmentManure                  1.7490820 2.5325847
TreatmentStopped                 0.4824153 1.2659180
After1970After                  -0.4356288 0.5239621
TreatmentFertilised:After1970After 0.3389668 1.6960332
TreatmentManure:After1970After   1.9356334 3.2926999
TreatmentStopped:After1970After -0.8726999 0.4843666
```

Mean of control group before 1970

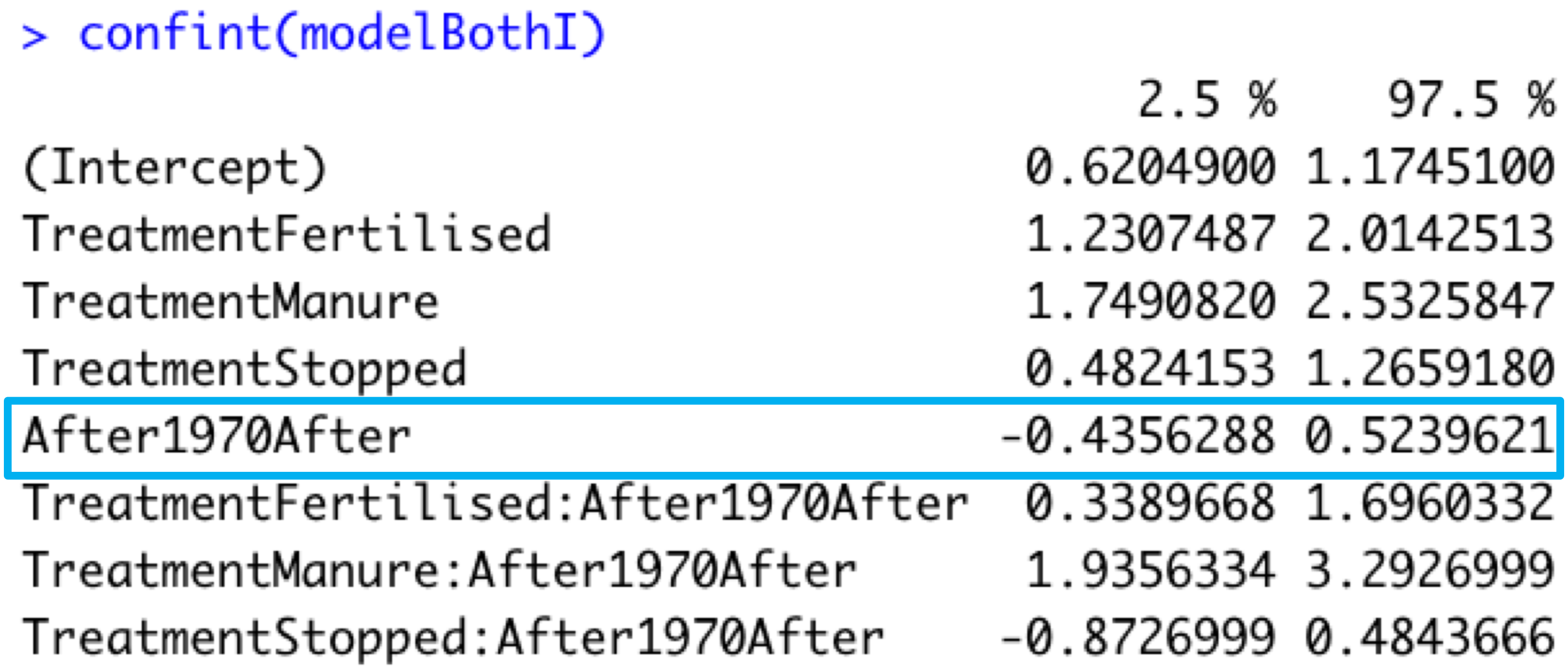
ANSWERS PART C



```
> confint(modelBothI)
                                       2.5 %    97.5 %
(Intercept)                        0.6204900 1.1745100
TreatmentFertilised                1.2307487 2.0142513
TreatmentManure                    1.7490820 2.5325847
TreatmentStopped                   0.4824153 1.2659180
After1970After                    -0.4356288 0.5239621
TreatmentFertilised:After1970After 0.3389668 1.6960332
TreatmentManure:After1970After     1.9356334 3.2926999
TreatmentStopped:After1970After   -0.8726999 0.4843666
```

Treatment effects – differences in mean caused by each treatment

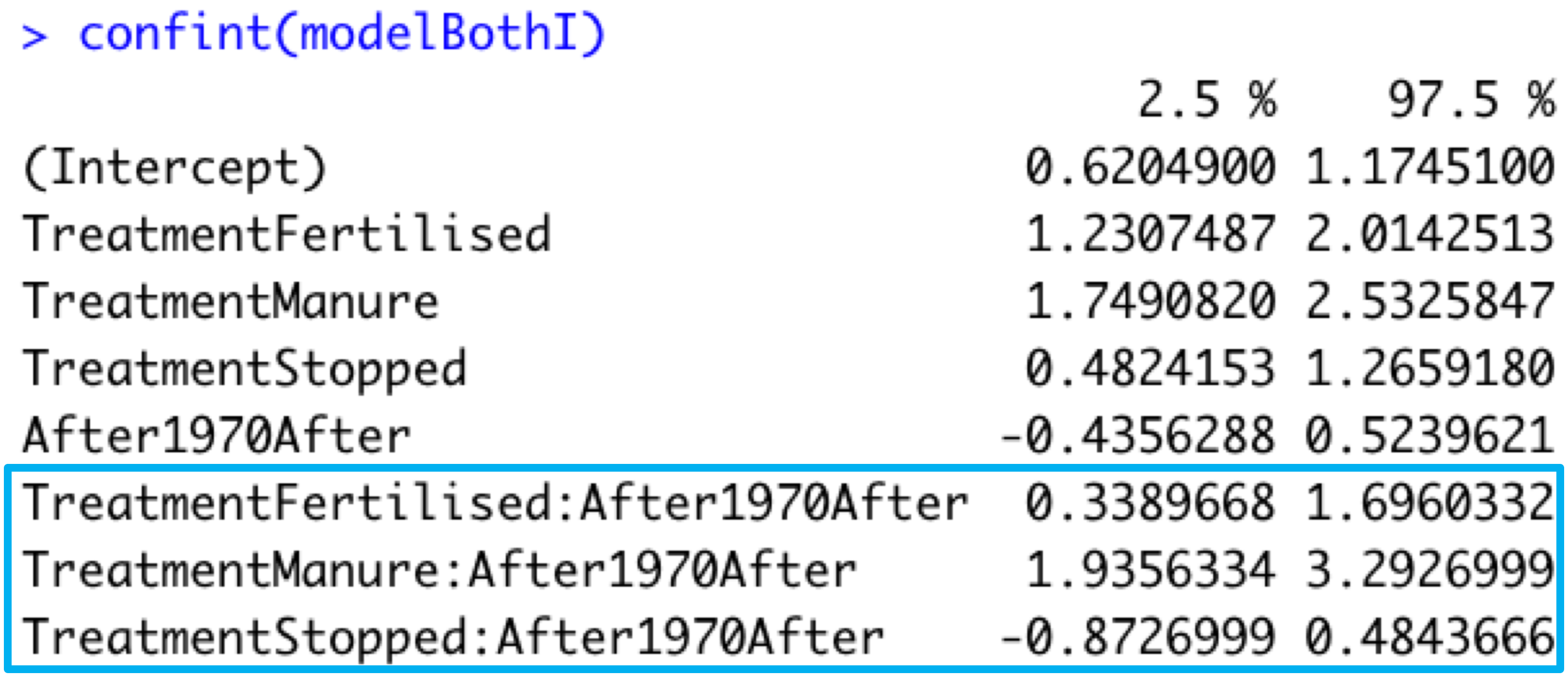
ANSWERS PART C



```
> confint(modelBothI)
                                     2.5 %    97.5 %
(Intercept)                      0.6204900 1.1745100
TreatmentFertilised              1.2307487 2.0142513
TreatmentManure                  1.7490820 2.5325847
TreatmentStopped                 0.4824153 1.2659180
After1970After                  -0.4356288 0.5239621
TreatmentFertilised:After1970After 0.3389668 1.6960332
TreatmentManure:After1970After   1.9356334 3.2926999
TreatmentStopped:After1970After -0.8726999 0.4843666
```

Time effect – differences in mean caused by change in time

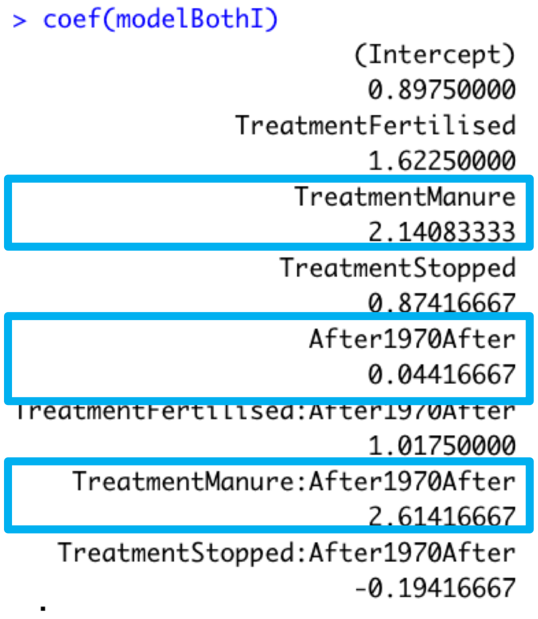
ANSWERS PART C



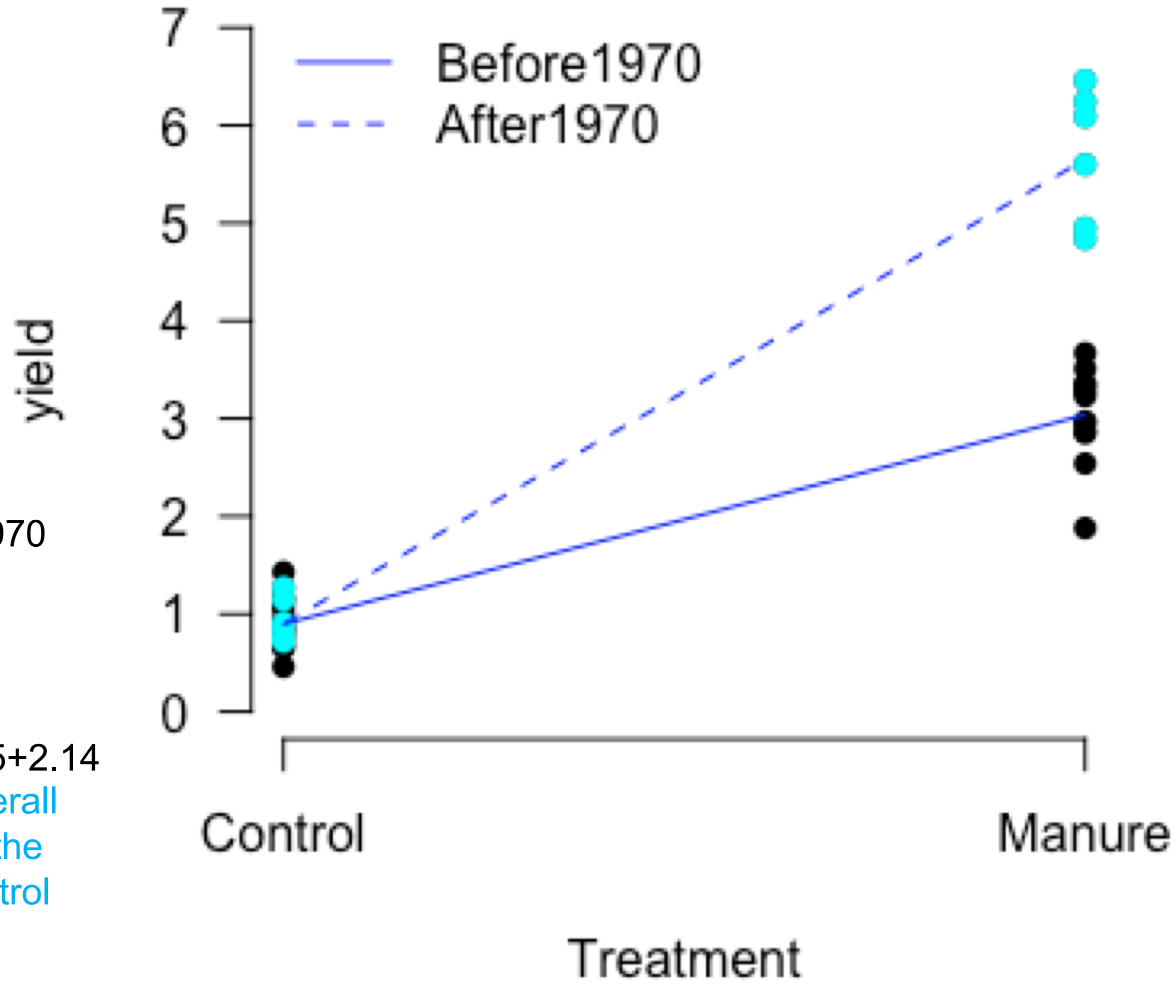
```
> confint(modelBothI)
                                     2.5 %    97.5 %
(Intercept)                      0.6204900 1.1745100
TreatmentFertilised              1.2307487 2.0142513
TreatmentManure                  1.7490820 2.5325847
TreatmentStopped                 0.4824153 1.2659180
After1970After                  -0.4356288 0.5239621
TreatmentFertilised:After1970After  0.3389668 1.6960332
TreatmentManure:After1970After   1.9356334 3.2926999
TreatmentStopped:After1970After -0.8726999 0.4843666
```

Interaction effects – differences in mean for each treatment from before 1970 to after 1970

ANSWERS PART C



```
> coef(modelBothI)
                       (Intercept)
                        0.89750000
              TreatmentFertilised
                        1.62250000
                   TreatmentManure
                        2.14083333
                  TreatmentStopped
                        0.87416667
                    After1970After
                        0.04416667
TreatmentFertilised:After1970After
                        1.01750000
    TreatmentManure:After1970After
                        2.61416667
   TreatmentStopped:After1970After
                       -0.19416667
 .
```



Intercept = mean of control before 1970

To get to manure before 1970 = 0.8975+2.14

To get to manure after 1970 = 0.8975+2.14 + 0.04 + 2.61 need to include the overall effect of time, then the interaction is the difference in time effect between control group and manure group

Categorical and continuous

REMEMBER

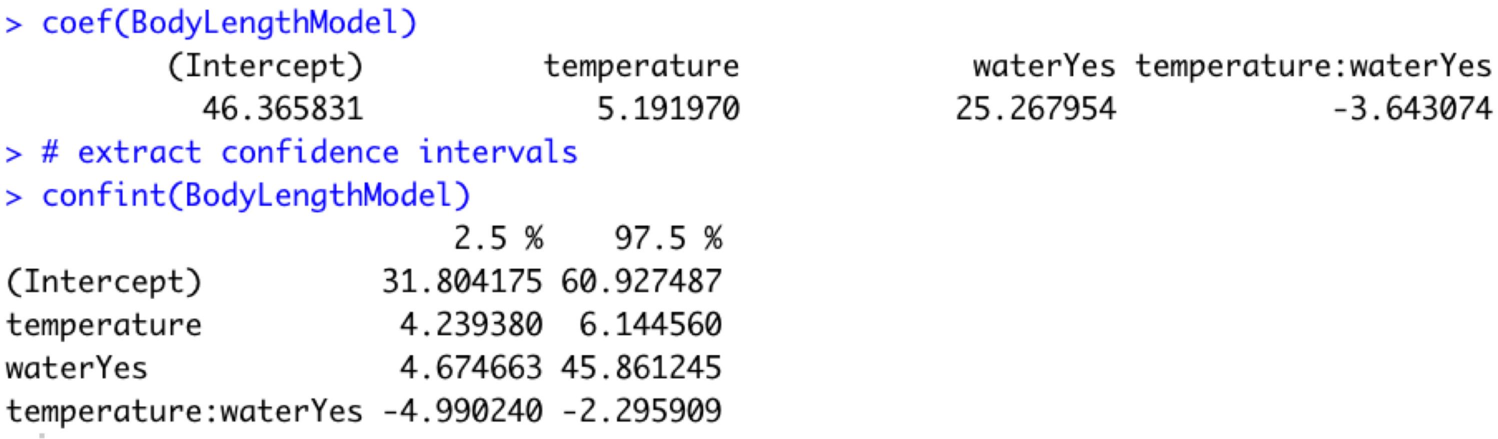
Categorical = in groups

Continuous = every value can exist

Exercise 4: Mixed continuous and categorical

- Start Part D of the module

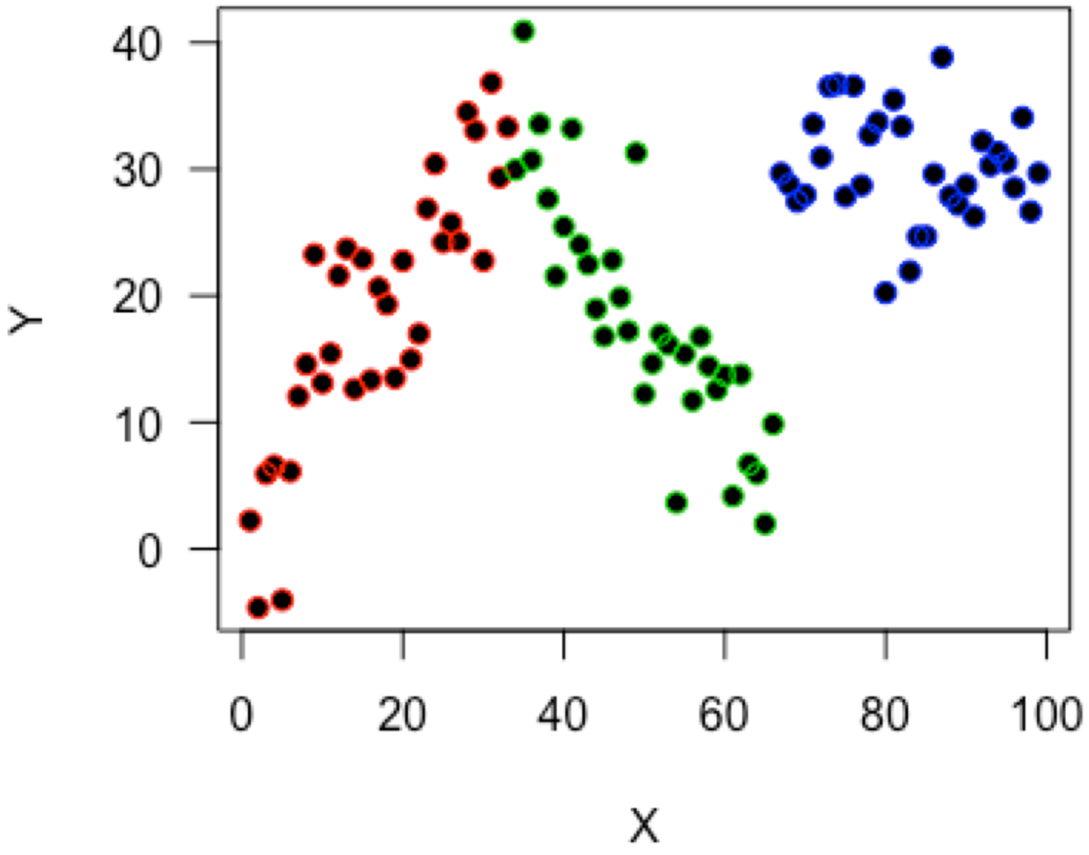
ANSWERS PART D1



```
> coef(BodyLengthModel)
          (Intercept)           temperature              waterYes temperature:waterYes 
            46.365831              5.191970             25.267954            -3.643074 
> # extract confidence intervals
> confint(BodyLengthModel)
                          2.5 %    97.5 %
(Intercept)           31.804175 60.927487
temperature            4.239380  6.144560
waterYes               4.674663 45.861245
temperature:waterYes  -4.990240 -2.295909
```

Interpreting!

Here we have both categorical and continuous variables



Categorical and continuous

Several ways we can model this

Y ~ X	Separately
Y ~ Groups	
Y ~ X + Groups	Additively
Y ~ X * Groups	Interaction

Categorical and continuous

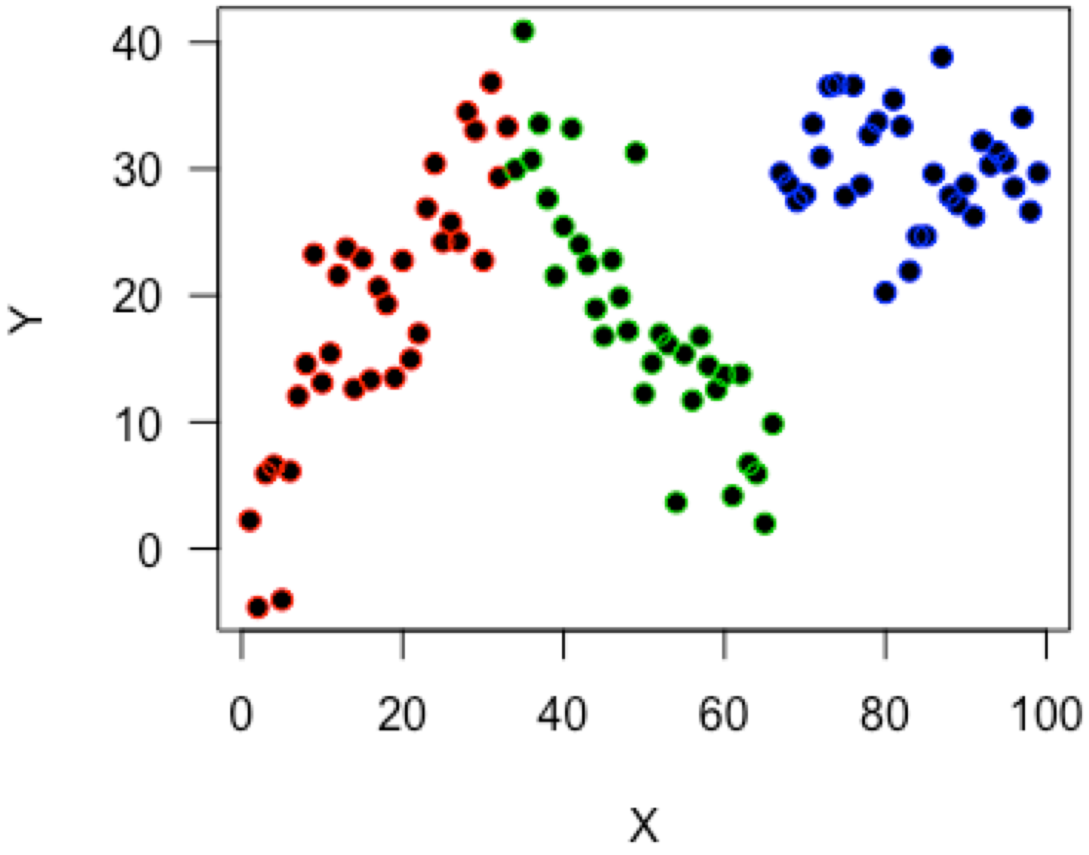
Several ways we can model this

Y ~ X	Separately
Y ~ Groups	
Y ~ X + Groups	Additively
Y ~ X * Groups	Interaction

Will depend on the effect of each

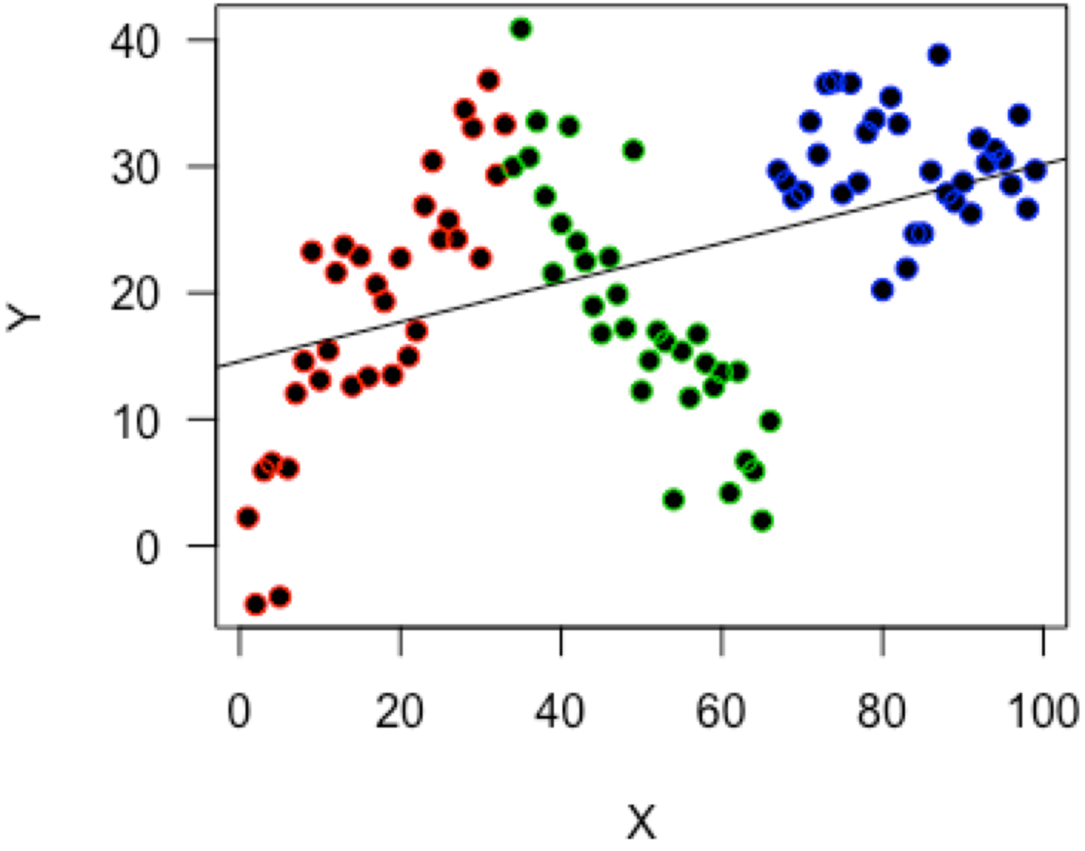
Categorical and continuous

Back to the example

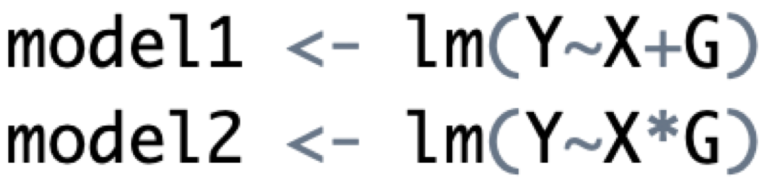


Categorical and continuous

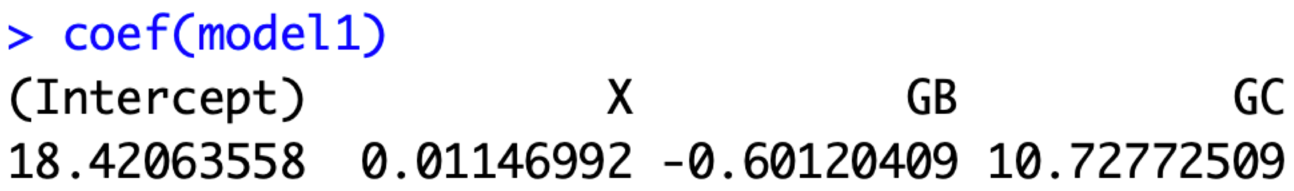
Back to the example



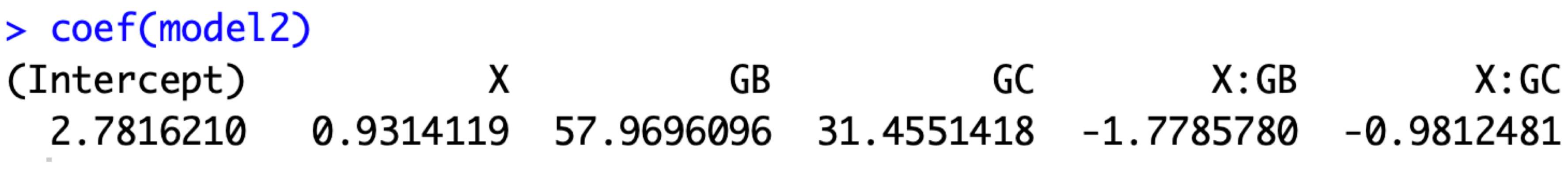
Interpreting



```
model1 <- lm(Y~X+G)
model2 <- lm(Y~X*G)
```



```
> coef(model1)
(Intercept)           X          GB          GC 
18.42063558  0.01146992 -0.60120409 10.72772509 
```

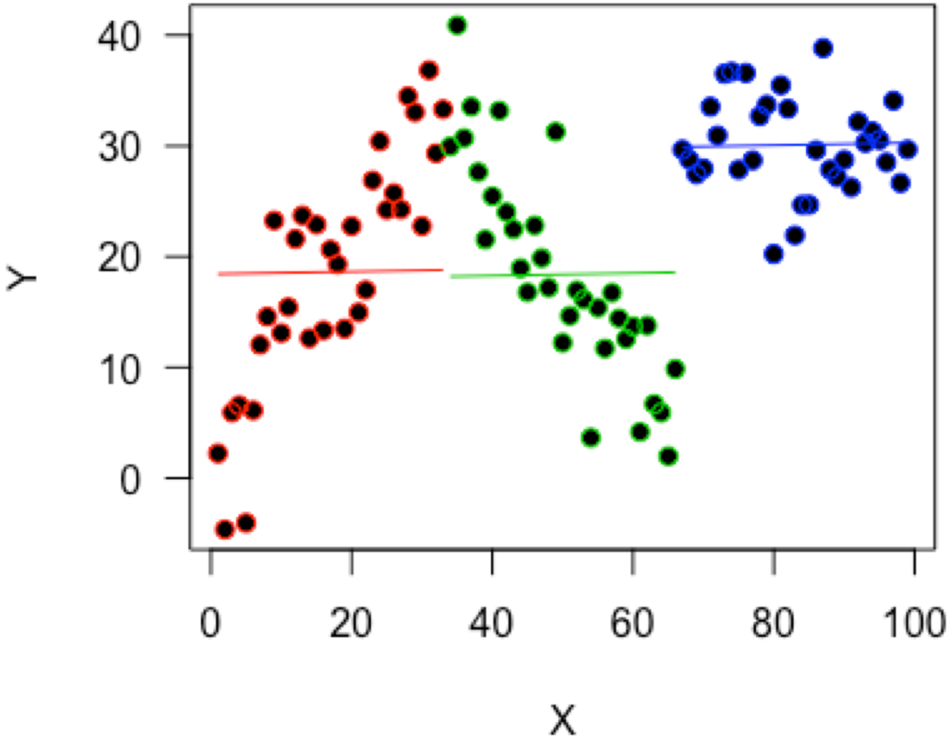


```
> coef(model2)
(Intercept)          X         GB         GC       X:GB       X:GC 
  2.7816210  0.9314119 57.9696096 31.4551418 -1.7785780 -0.9812481 
```

No interaction

```
model1 <- lm(Y~X+G)
```

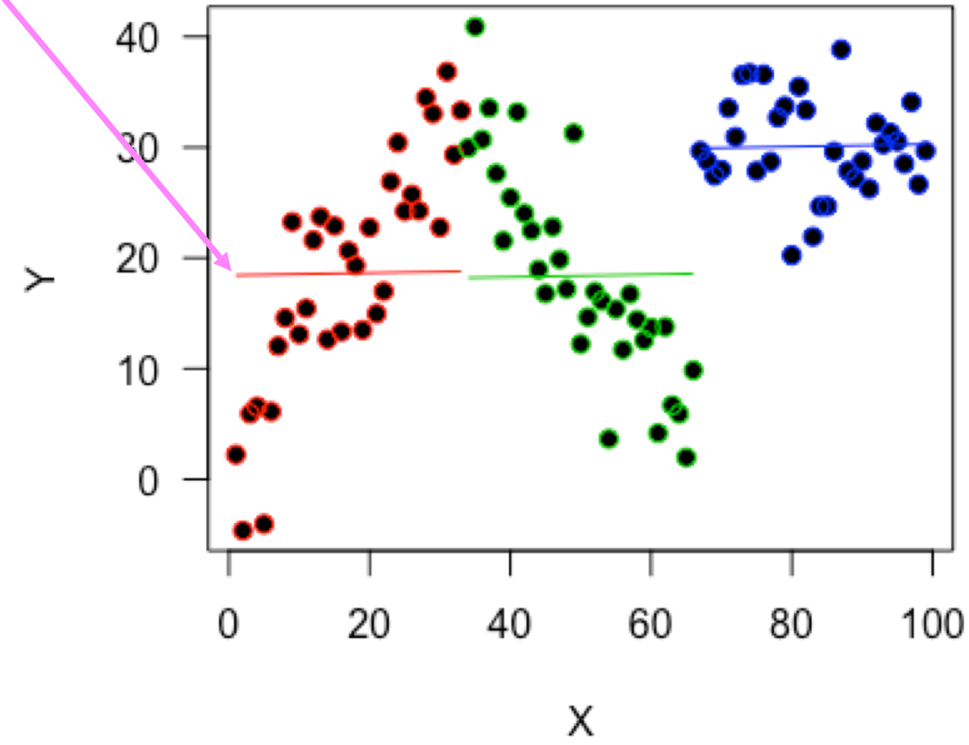
```
> coef(model1)
(Intercept)           X          GB          GC 
18.42063558  0.01146992 -0.60120409 10.72772509 
```



No interaction

```
model1 <- lm(Y~X+G)
```

```
> coef(model1)
(Intercept)           X          GB          GC 
18.42063558  0.01146992 -0.60120409 10.72772509 
```

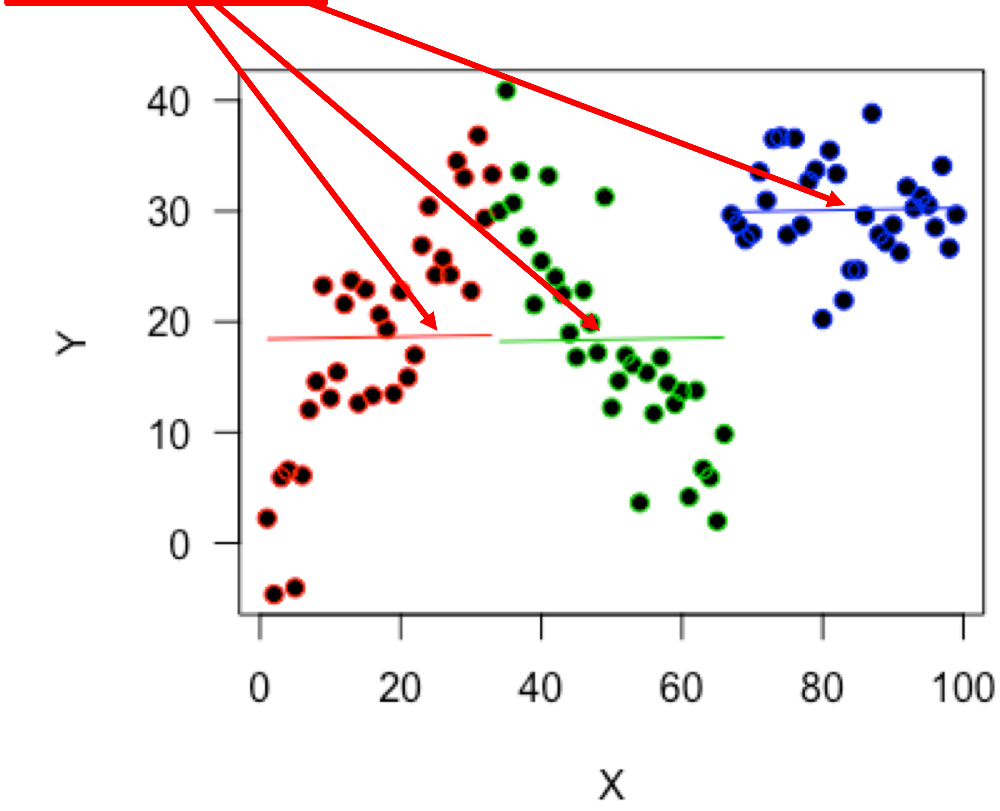


Intercept of line of Group A

No interaction

```
model1 <- lm(Y~X+G)
```

```
> coef(model1)
(Intercept)           X          GB          GC
18.42063558  0.01146992 -0.60120409 10.72772509
```



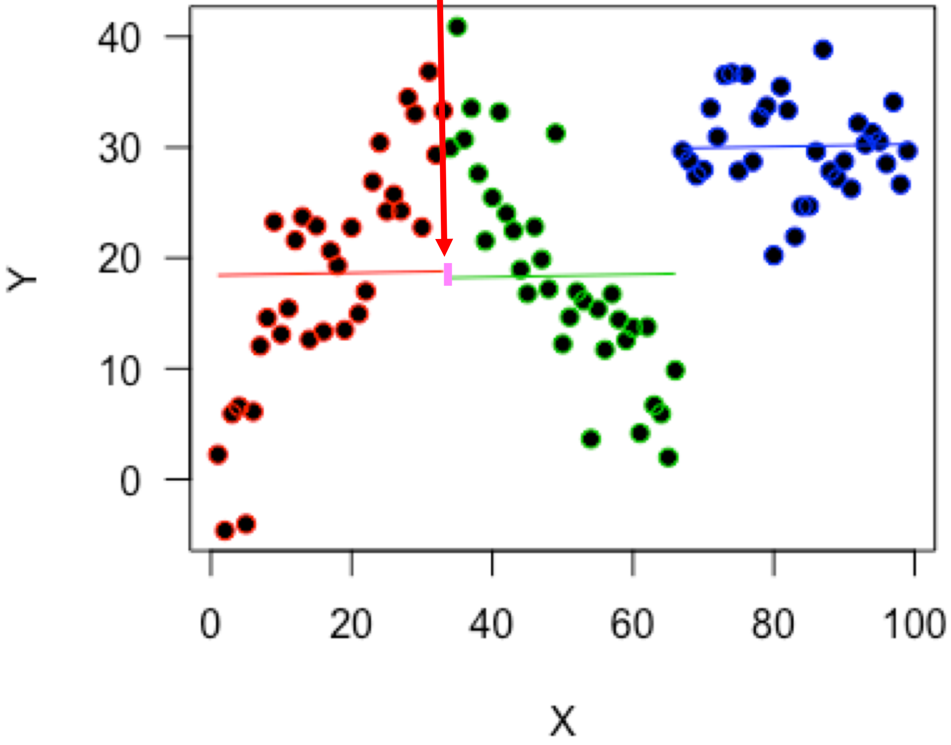
Slope value for all groups (same)

No interaction

```
model1 <- lm(Y~X+G)
```

```
> coef(model1)
(Intercept)           X          GB          GC
18.42063558  0.01146992 -0.60120409 10.72772509
```

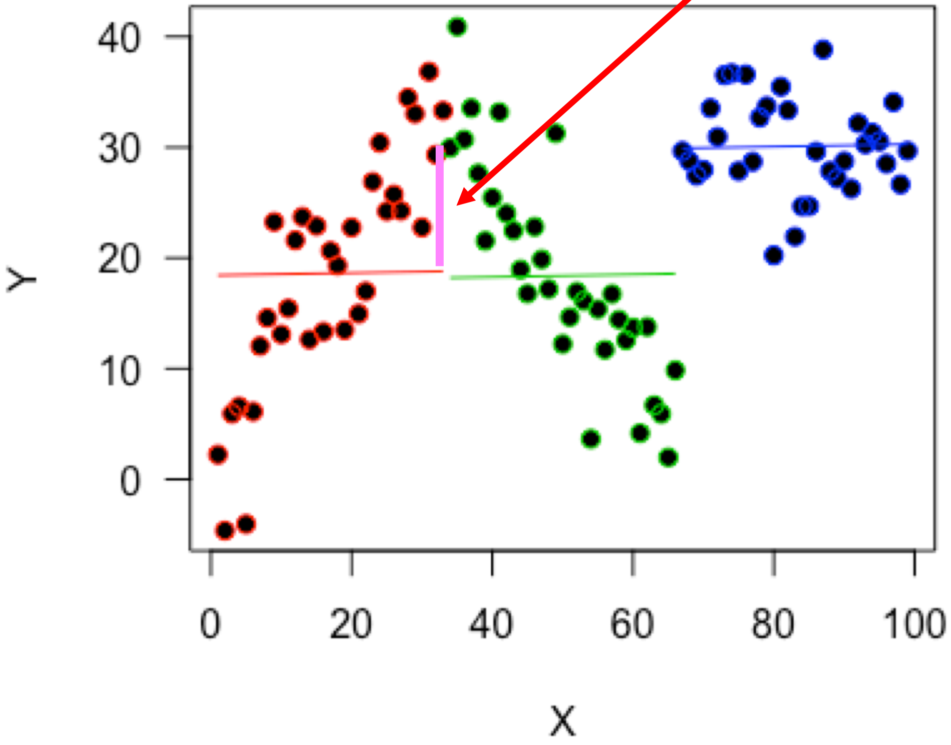
Difference in intercept from Group A to Group B



No interaction

```
model1 <- lm(Y~X+G)
```

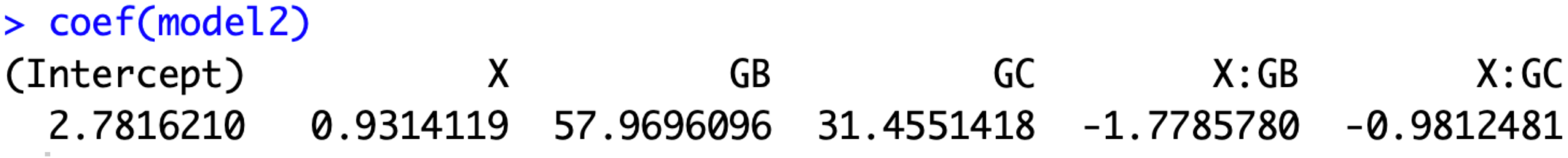
```
> coef(model1)
(Intercept)           X          GB          GC 
18.42063558  0.01146992 -0.60120409 10.72772509 
```



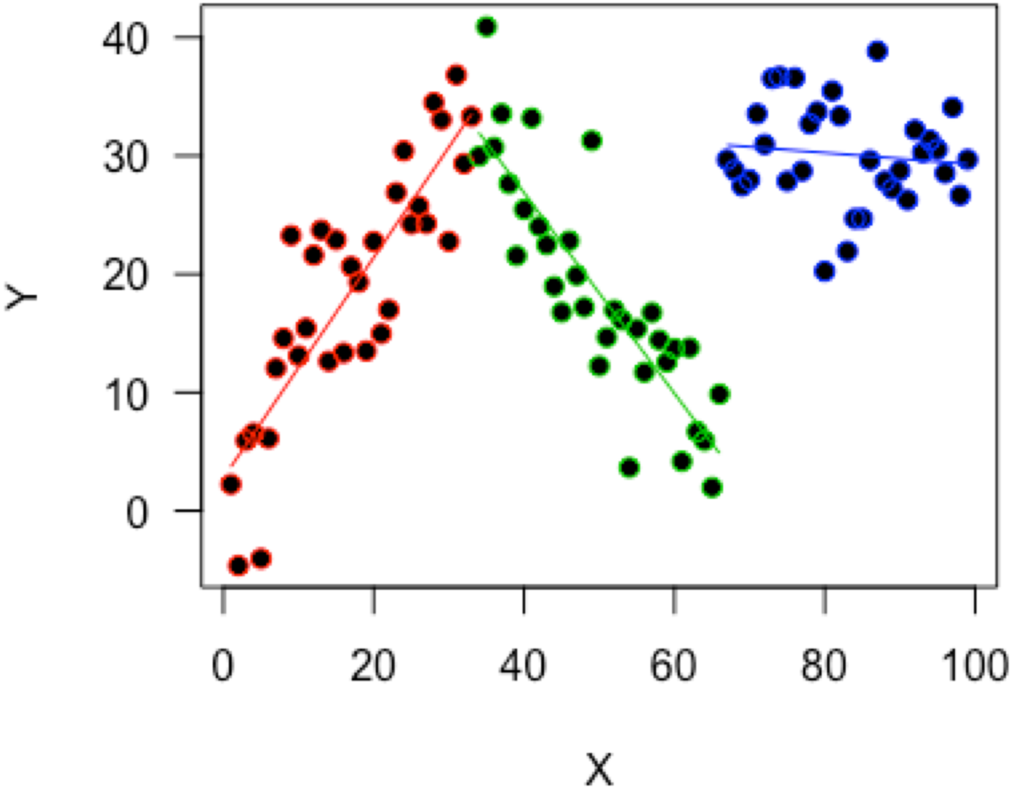
Difference in intercept from Group A to Group C

Interaction

```
model2 <- lm(Y~X*G)
```

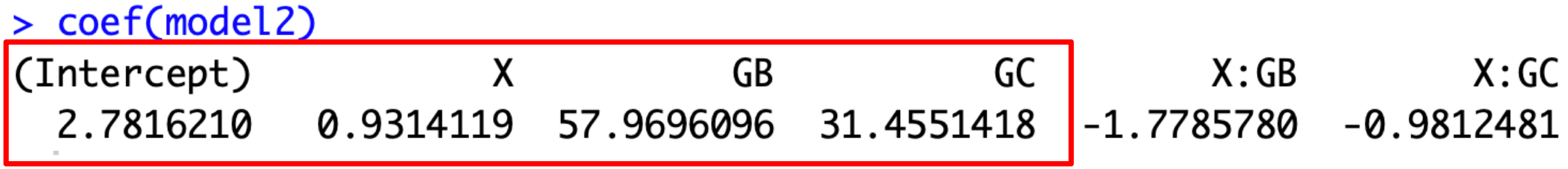


```
> coef(model2)
(Intercept)           X          GB          GC        X:GB        X:GC 
  2.7816210   0.9314119  57.9696096  31.4551418  -1.7785780  -0.9812481 
```

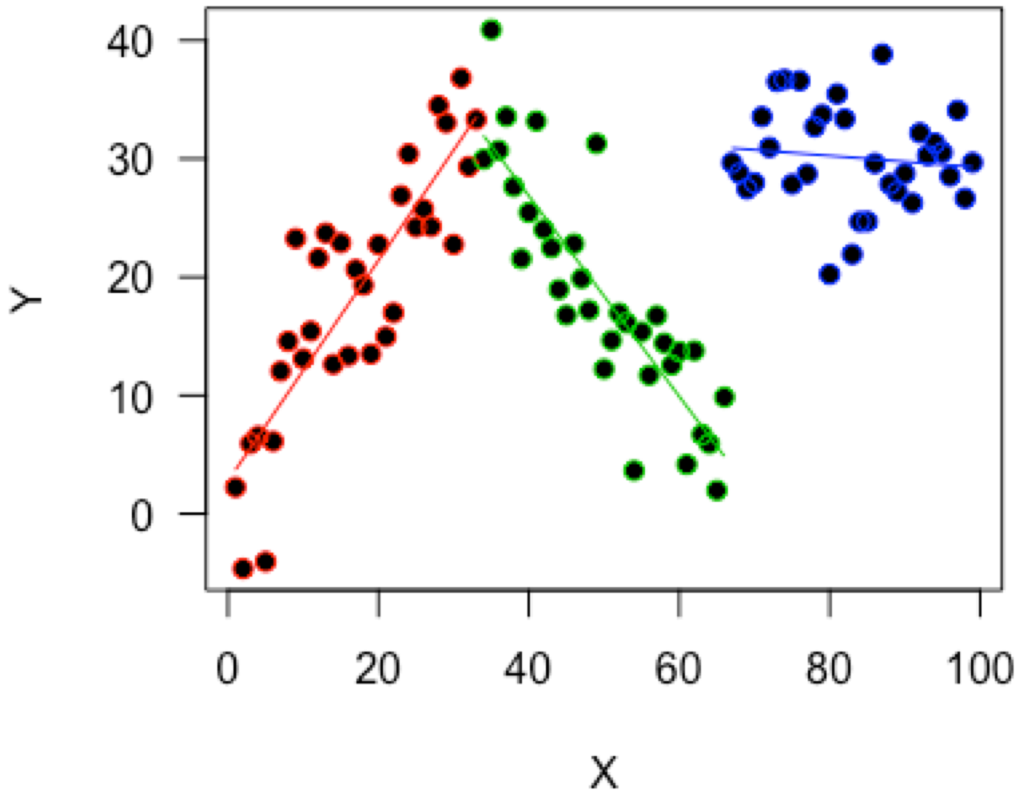


Interaction

```
model2 <- lm(Y~X*G)
```

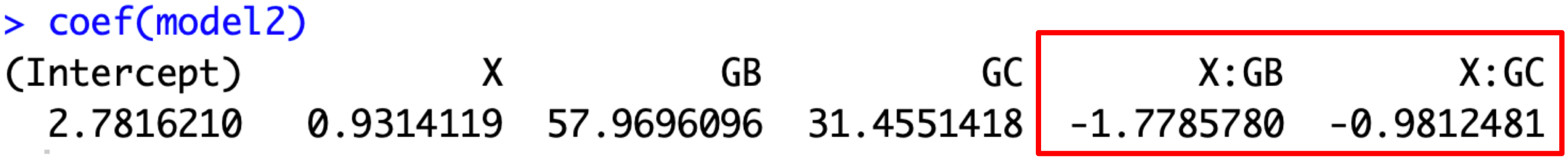


```
> coef(model2)
(Intercept)           X          GB          GC        X:GB        X:GC 
  2.7816210   0.9314119  57.9696096  31.4551418  -1.7785780  -0.9812481 
```



Interaction

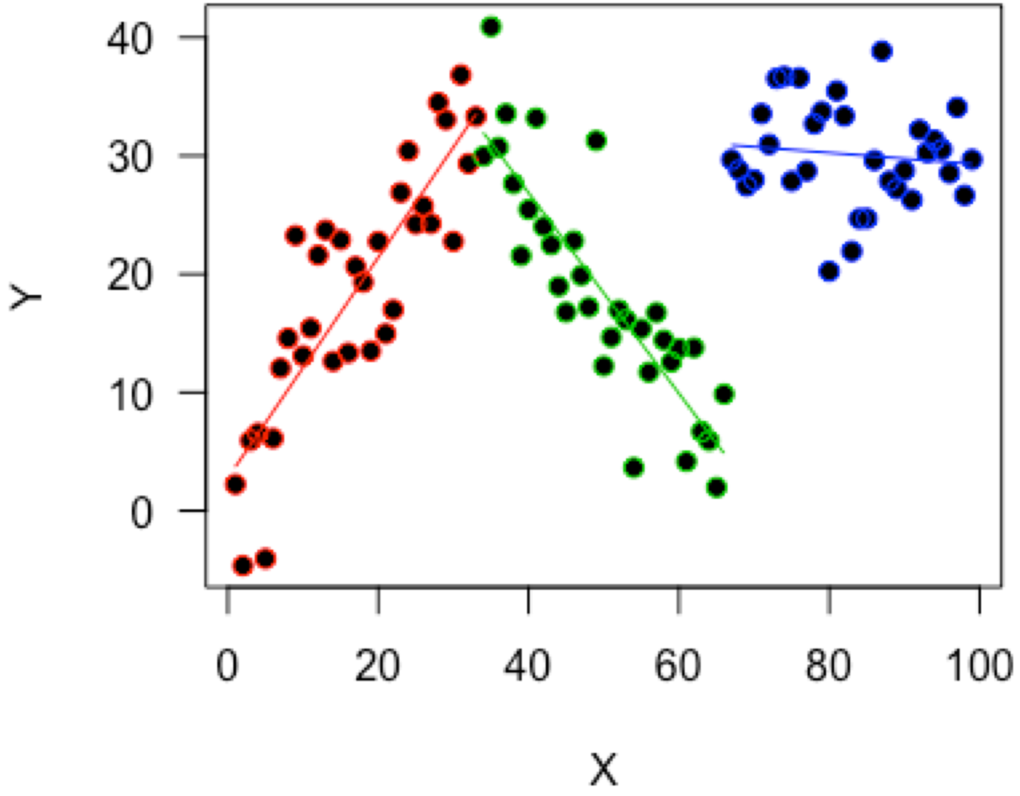
```
model2 <- lm(Y~X*G)
```



```
> coef(model2)
(Intercept)           X          GB          GC        X:GB        X:GC 
  2.7816210   0.9314119  57.9696096  31.4551418  -1.7785780  -0.9812481 
```

Differences in slopes

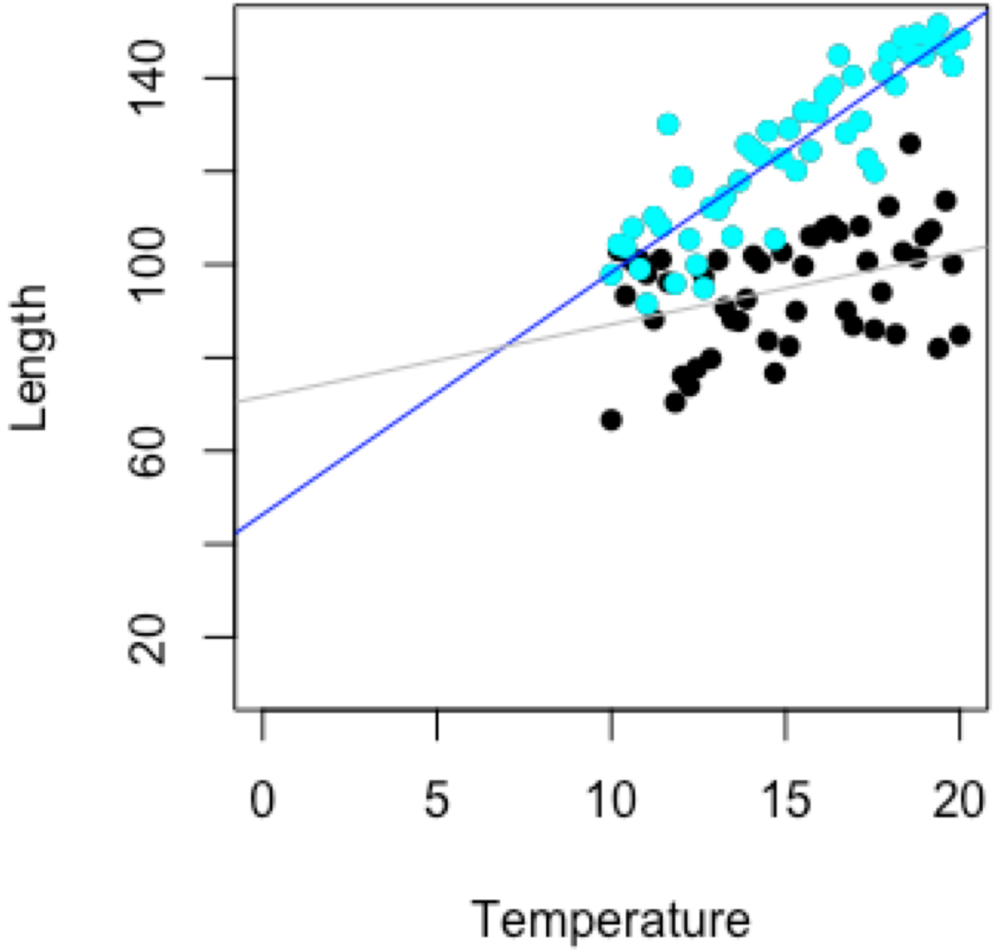
Interaction!



Exercise 4: Mixed continuous and categorical

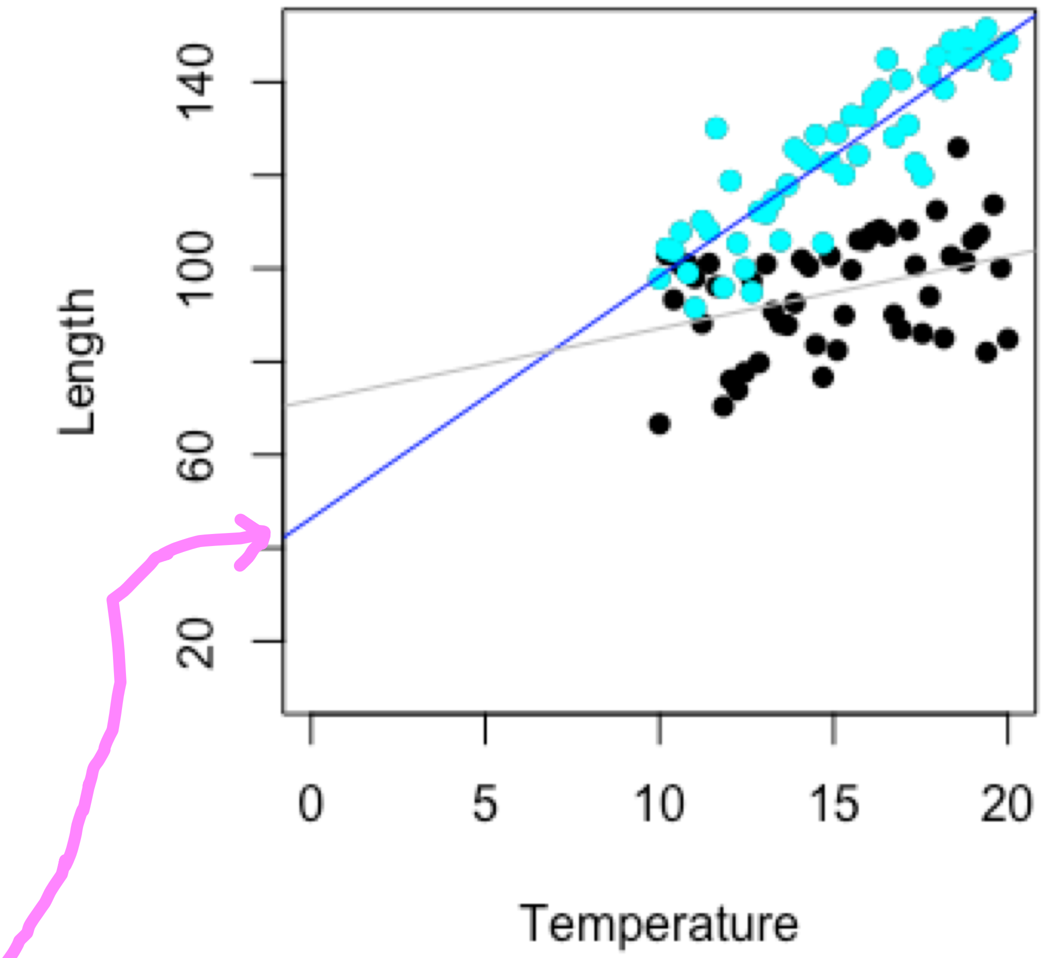
- Complete Part D of the module

ANSWERS PART D2



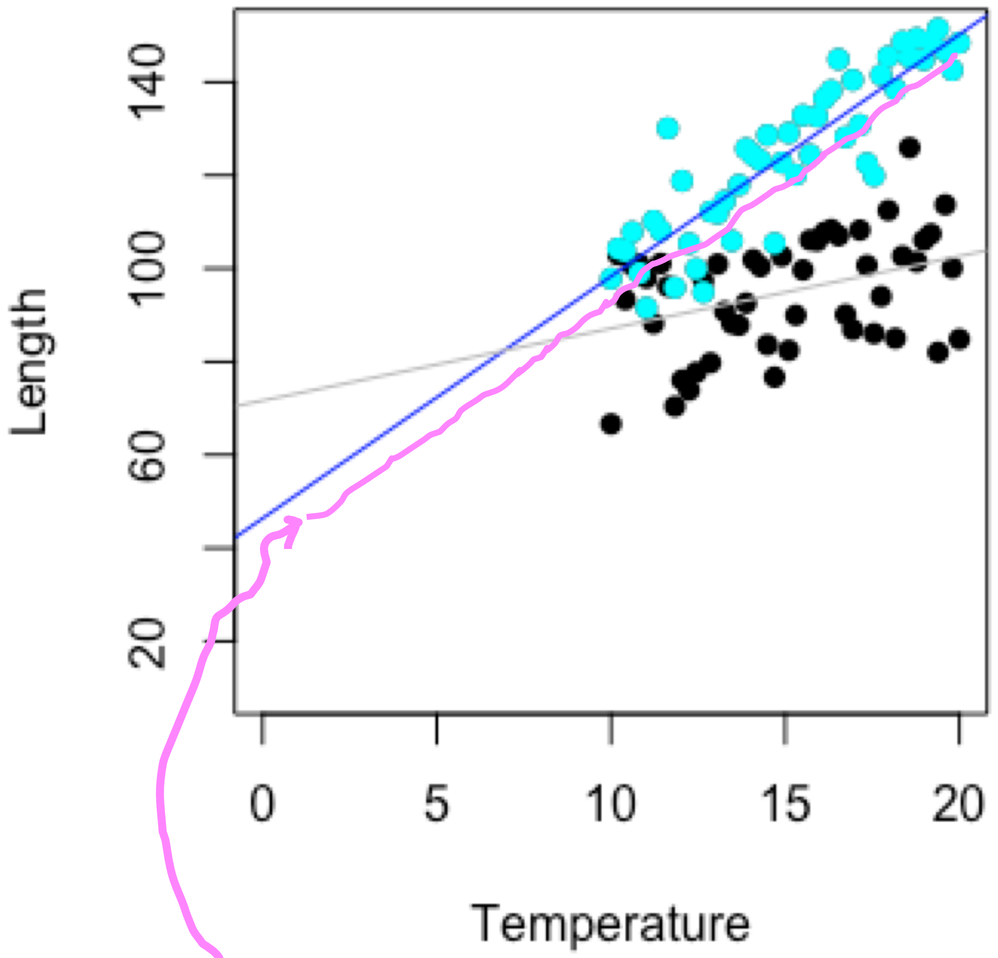
```
> coef(BodyLengthModel)
          (Intercept)           temperature              waterYes temperature:waterYes
            46.365831              5.191970             25.267954            -3.643074
```

ANSWERS PART D2



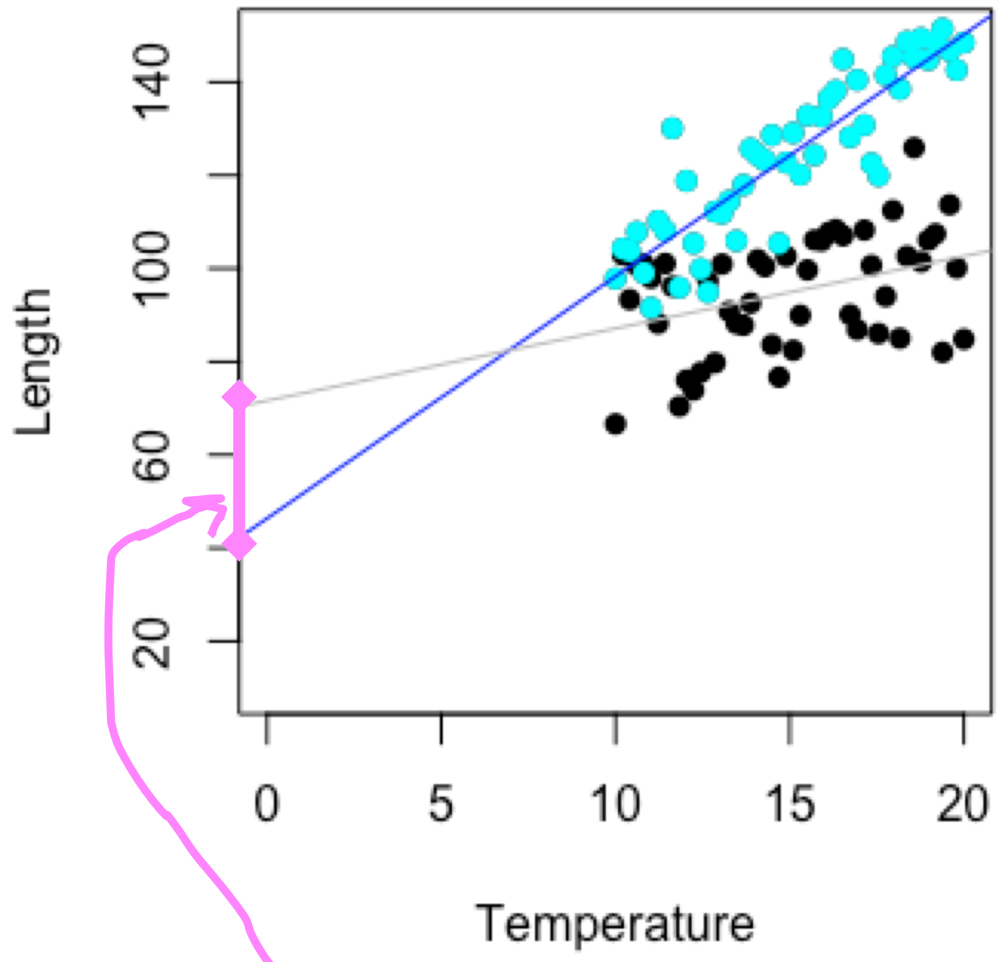
```
> coef(BodyLengthModel)
         (Intercept)          temperature             waterYes temperature:waterYes
           46.365831             5.191970            25.267954            -3.643074
```

ANSWERS PART D2



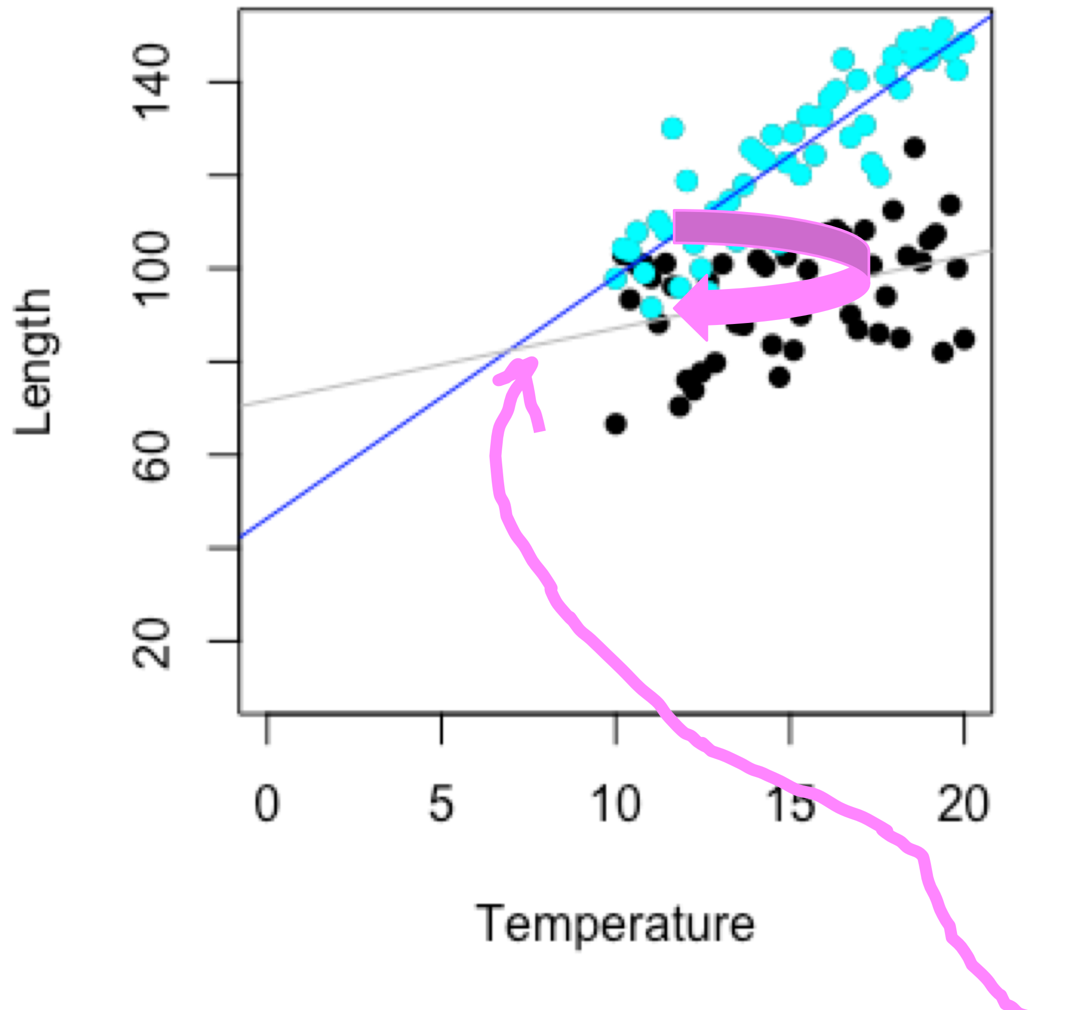
```
> coef(BodyLengthModel)
         (Intercept)          temperature             waterYes temperature:waterYes
           46.365831             5.191970            25.267954            -3.643074
```

ANSWERS PART D2



```
> coef(BodyLengthModel)
         (Intercept)          temperature             waterYes temperature:waterYes
           46.365831             5.191970            25.267954            -3.643074
```

ANSWERS PART D2



```
> coef(BodyLengthModel)
        (Intercept)          temperature                waterYes temperature:waterYes
          46.365831             5.191970               25.267954            -3.643074
```

ANSWERS PART D2

Temperature has positive effect on body length (warmer = longer)

The strength of that effect is bigger when there is no water

But the effect of water itself, is to increase body length

Large uncertainty in the effect of water, but still doesn’t cross 0

Does seem to be interaction

Summary

When we combine categorical and continuous explanatory variables….

Drawing several lines – one per group

No interaction = different intercepts

Interaction = different intercepts and slopes

All about lines!

Tips and tricks to reading output

What went in?

Sometimes you will be given output and won’t know what went in

OR you might need to check that what you put in is behaving how you expect

How can we tell how R is treating our variables?

Tip 1: Look at your data

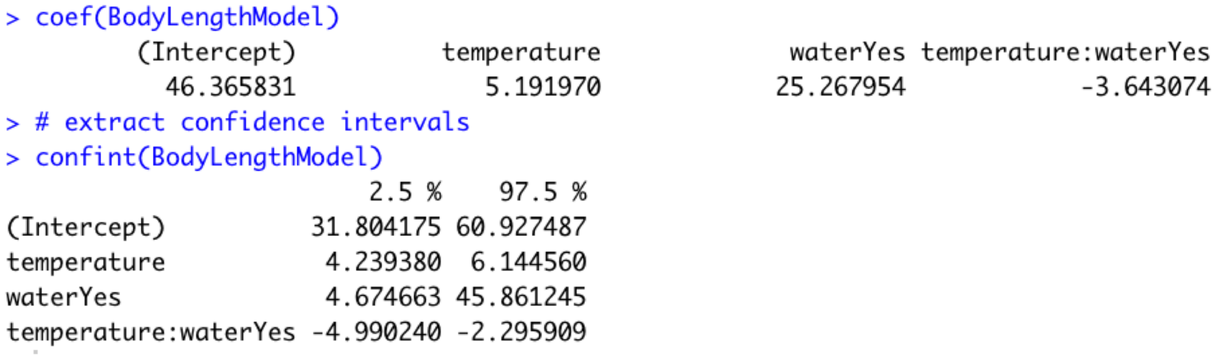
Read the data description

Look the data if possible

Ask: is it categorical or continuous?

Tip 2: Look at the beta value in output

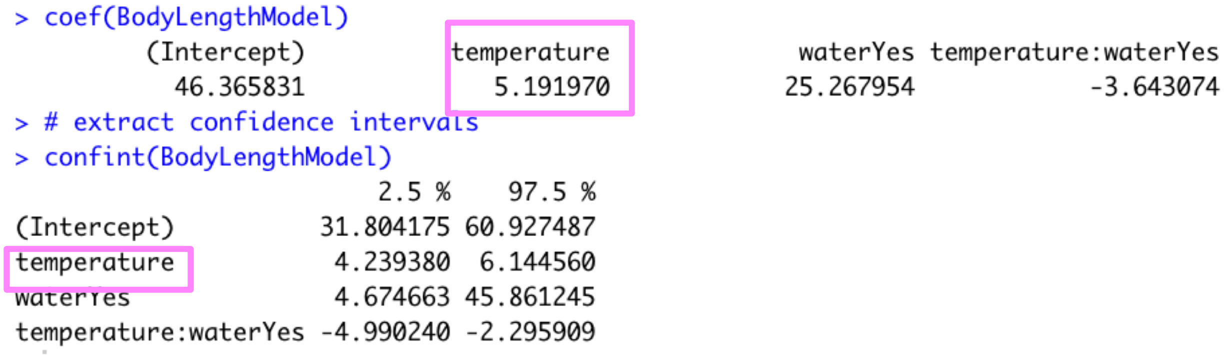
Is it just the variable name? Or anything else there?



```
> coef(BodyLengthModel)
         (Intercept)          temperature             waterYes temperature:waterYes 
           46.365831             5.191970            25.267954            -3.643074 
> # extract confidence intervals
> confint(BodyLengthModel)
                          2.5 %    97.5 %
(Intercept)           31.804175 60.927487
temperature            4.239380  6.144560
waterYes               4.674663 45.861245
temperature:waterYes  -4.990240 -2.295909
```

Tip 2: Look at the beta value in output

Is it just the variable name? Or anything else there?

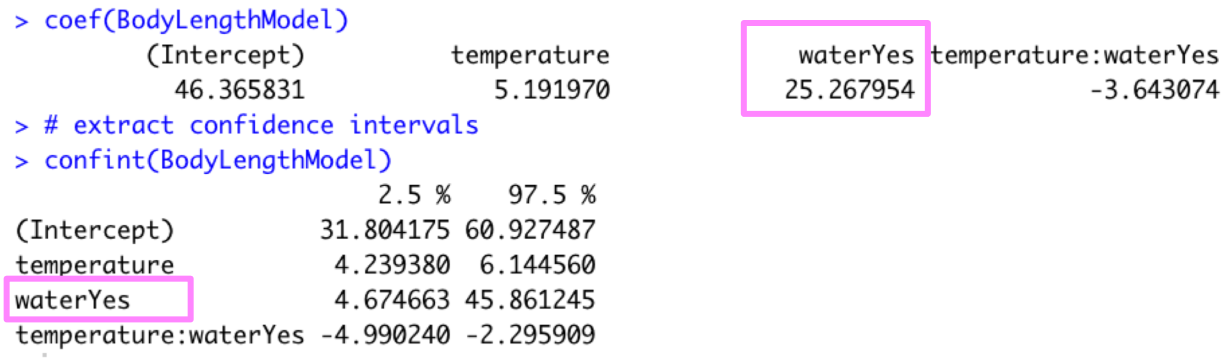


```
> coef(BodyLengthModel)
         (Intercept)          temperature             waterYes temperature:waterYes 
           46.365831             5.191970            25.267954            -3.643074 
> # extract confidence intervals
> confint(BodyLengthModel)
                          2.5 %    97.5 %
(Intercept)           31.804175 60.927487
temperature            4.239380  6.144560
waterYes               4.674663 45.861245
temperature:waterYes  -4.990240 -2.295909
```

Variable name only = continuous

Tip 2: Look at the beta value in output

Is it just the variable name? Or anything else there?

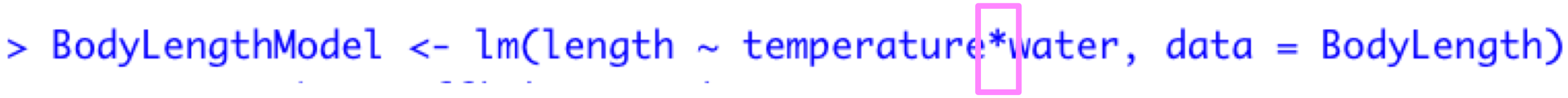


```
> coef(BodyLengthModel)
         (Intercept)          temperature             waterYes temperature:waterYes
           46.365831             5.191970            25.267954            -3.643074
> # extract confidence intervals
> confint(BodyLengthModel)
                         2.5 %    97.5 %
(Intercept)          31.804175 60.927487
temperature           4.239380  6.144560
waterYes              4.674663 45.861245
temperature:waterYes -4.990240 -2.295909
```

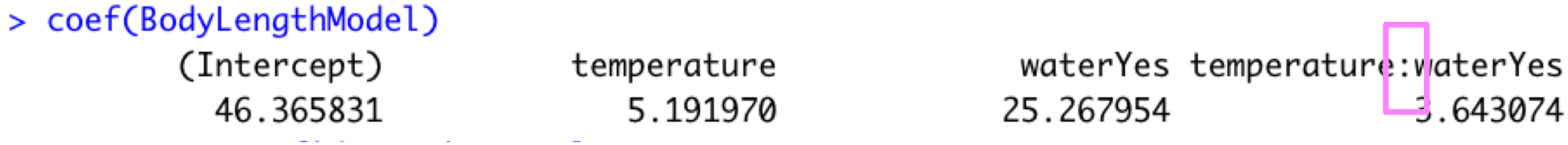
Group name too = categorical

Tip 3: Look for * and :

You can see when an interaction is included



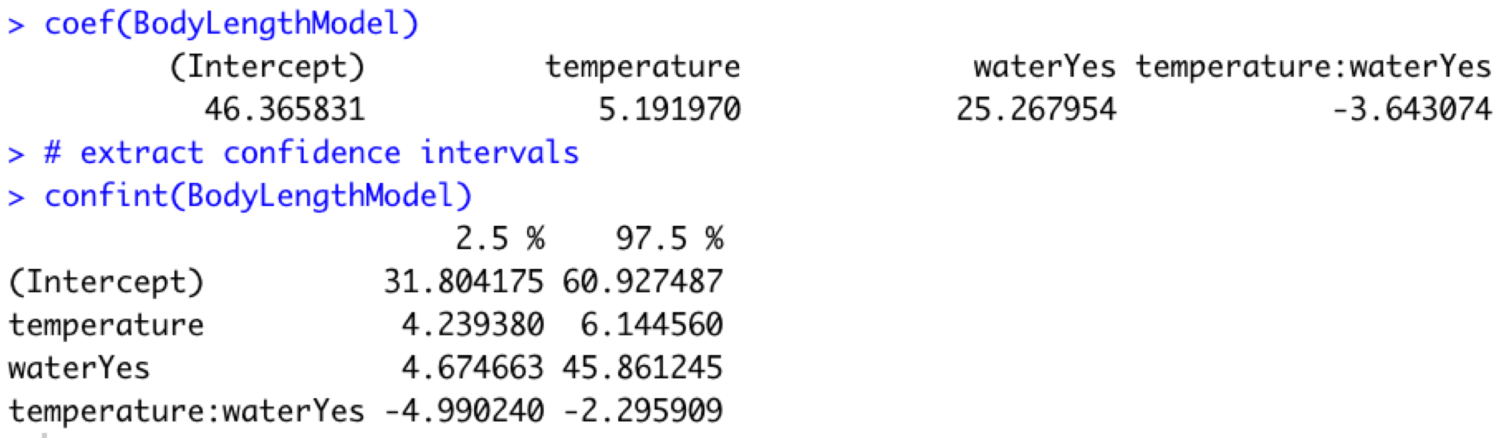
```
> BodyLengthModel <- lm(length ~ temperature*water, data = BodyLength)
```



```
> coef(BodyLengthModel)
          (Intercept)           temperature              waterYes temperature:waterYes
            46.365831              5.191970             25.267954             3.643074
```

Tip 4: What is missing?

Tells you the intercept



```
> coef(BodyLengthModel)
         (Intercept)          temperature             waterYes temperature:waterYes 
           46.365831             5.191970            25.267954            -3.643074 
> # extract confidence intervals
> confint(BodyLengthModel)
                          2.5 %    97.5 %
(Intercept)           31.804175 60.927487
temperature            4.239380  6.144560
waterYes               4.674663 45.861245
temperature:waterYes  -4.990240 -2.295909
```

Water = No is missing here

Tip 5: Remember what went in

If continuous went in, will expect a continuous line

If it did not, differences in means

Exercise 5: Detective skills

- Complete Part E of the module

Summary

Recap of last week

- EX1: How to choose a model

More than one categorical variable

- EX2: Two categorical variables
- EX3: Interactions

Mixing categorical and continuous

- EX4: Categorical and continuous

Tips and tricks to reading outputs

- EX5: What has been done?

Tomorrow

I need to go to teaching seminar

Exam style practice – mark scheme online so can practice grading yourselves

Email me if any things not clear so far – can maybe do 10 mins on it next week