Comparing more than two numeric variables: Multiple linear regression

Research Methods for Human Inquiry

Andrew Perfors

Our analysis last time suggested that having a range of species increased the productivity of the land!

Hang on a moment, though. What if it's just that bigger land is more productive, and bigger land has a wider range of species?

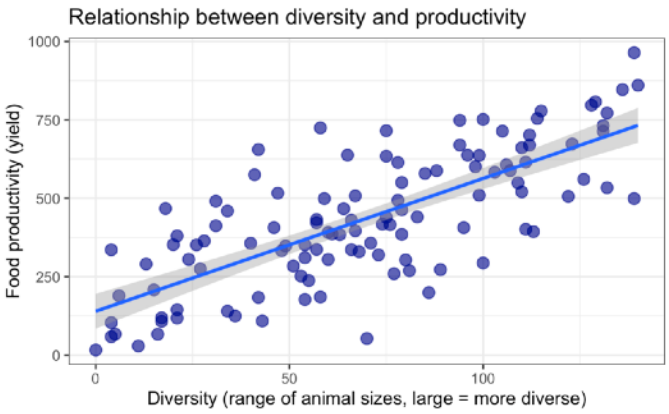
Oooh, yeah.
That's a problem.
How can we tell?

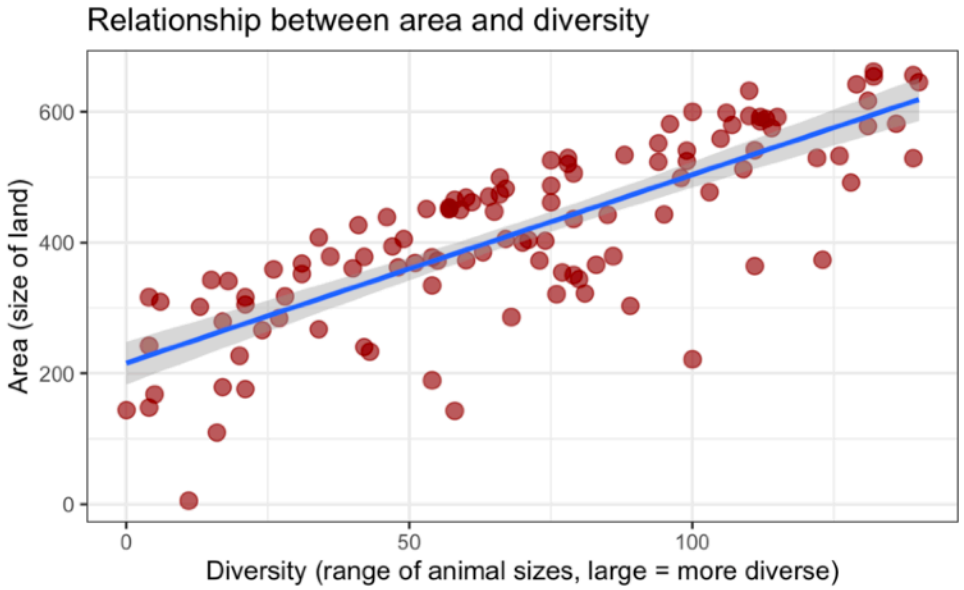
Remember our dataset

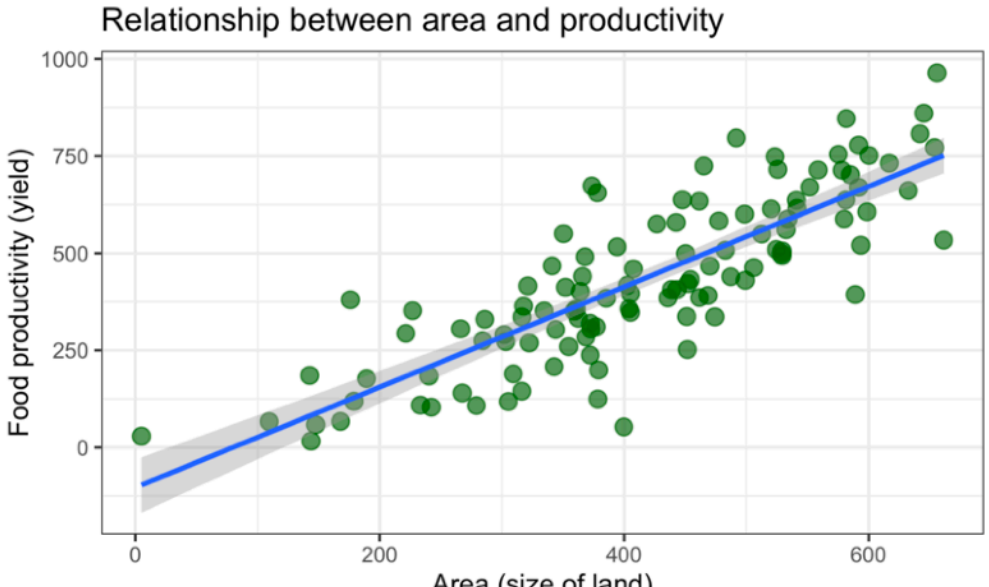
- One tibble, `d`, with four variables
  - `land`... code uniquely identifying each plot of land
  - `area`... size of each plot of land
  - `diversity`... measure of diversity of species farming each land; it's the maximum sized person minus the minimum sized person (so larger is more diverse)
  - `yield`... units of each food this plot of land yielded

```
> head(d)
      land    area range yield
1 iUxbR6F2 578.175   131 713.6
2 1iDku4Bd 491.714   128 796.7
3 4wTCpKQy 285.915    68 329.7
4 cj47Q9GZ 443.127    95 406.4
5 hnnu0h26 189.164    54 177.2
6 zvbOpHaV 368.642    51 284.0
```

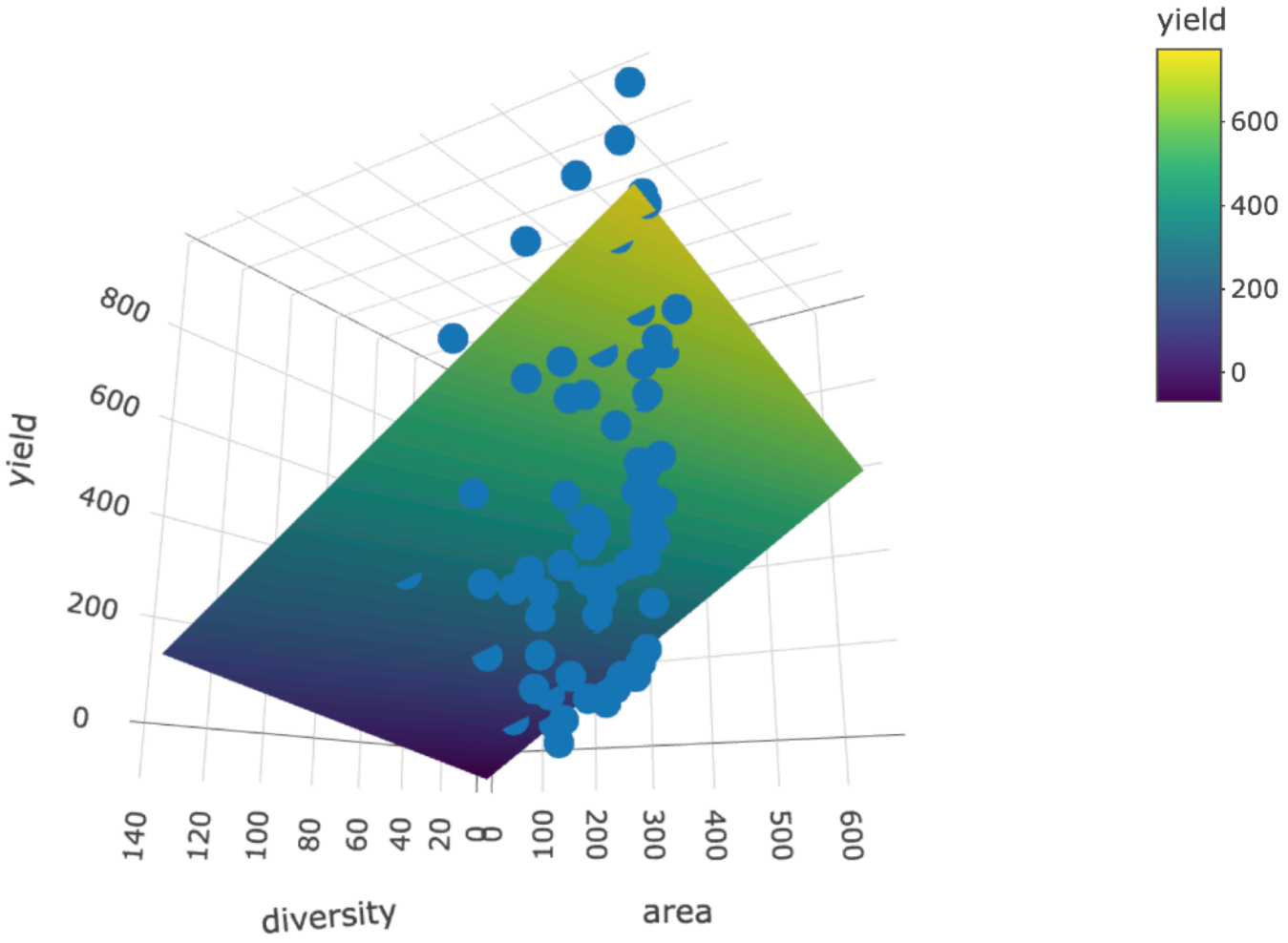
First, we can look at all scatterplots







But this is 3D so let's look at the 3D version



week9day2plots.html (First plot on there, model without interaction) Go to the html to rotate it!

Multiple regression

- So far, we've only tried to explain the outcome in terms of a single predictor.

  1 predictor = line

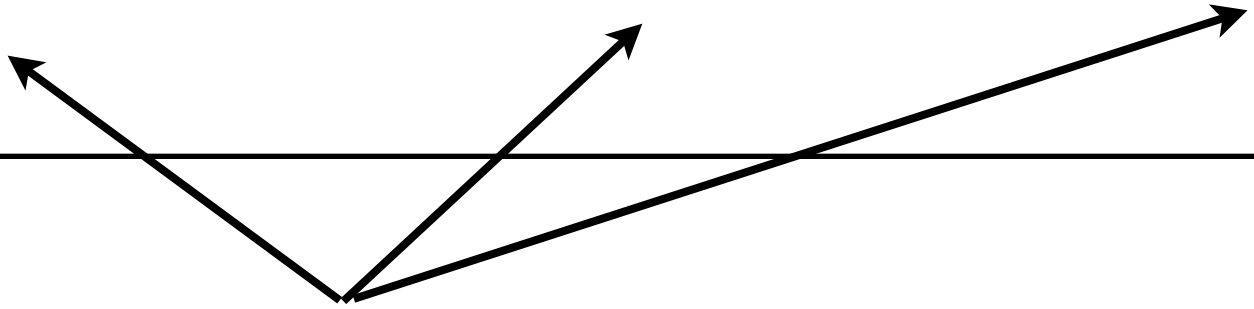
  - "Simple linear regression"
  - Mostly the same thing as Pearson correlations

- In real life, we usually expect that multiple variables could predict our outcome variable
  - "Multiple linear regression"

  2 predictors = plane

A multiple regression

$$\text{Yield} = b_1 * \text{Diversity} + b_2 * \text{Area} + b_0$$



Consider now *two* slope terms and one intercept...
these are called "regression coefficients"

A multiple regression

$$\text{Yield} = b_1 * \text{Diversity} + b_2 * \text{Area} + b_0$$

$$Y_i = b_1 X_{1i} + b_2 X_{2i} + b_0 + \epsilon_i$$

The regression model, using variables instead of words

```
yield ~ diversity + area
```

The R formula that describes this model

Let's run the model

```
> model2 <- lm(yield ~ diversity + area, data=d)
> model2
```

The formula uses an outcome variable of `yield` and *two* predictors, `diversity` and `area` (i.e., how much is food yield of a plot of land affected by the diversity of species farming the land and the area of the land?)

The dataset is called `d`

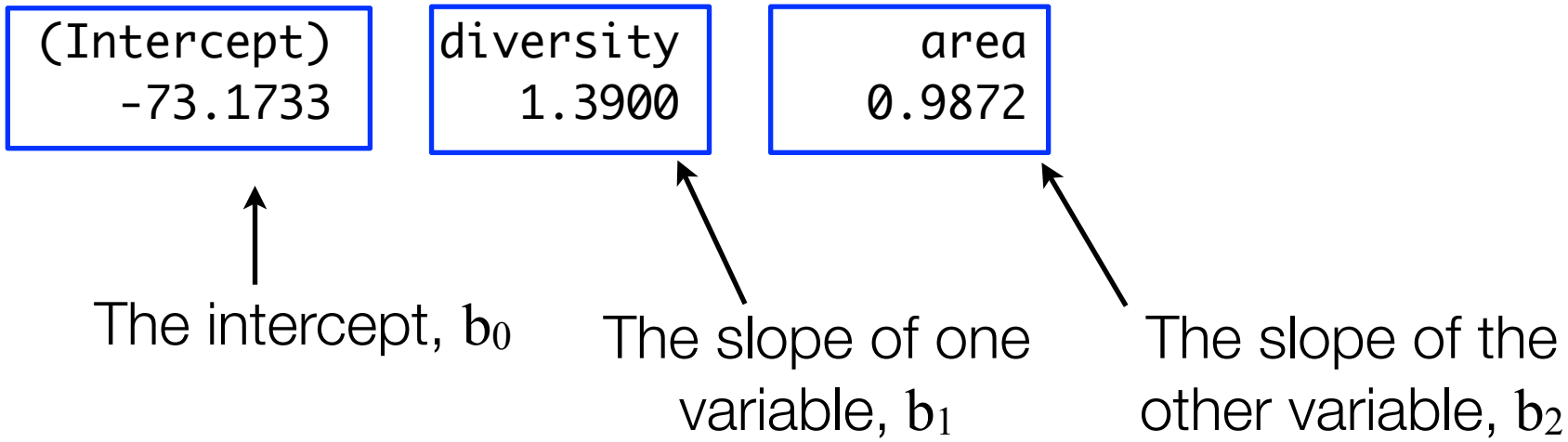
This command is similar, but the formula is different and the results are stored in a variable called `model2`

Let's run the model

```
> model2 <- lm(yield ~ diversity + area, data=d)
> model2

Call:
lm(formula = yield ~ diversity + area, data = d)

Coefficients:
(Intercept)    diversity         area
   -73.1733       1.3900       0.9872
```



Let's run the model

Notice that adding another variable changes everything!

This is for a similar reason that adding variables to an ANOVA changes things: additional variables explain some of the variance. In a multiple regression with two variables, you're trying to find the plane that best fits *both of them*. The interpretation of the coefficients changes accordingly.

```
> model1 <- lm(yield ~ diversity, data=d )
(Intercept)    diversity
    139.877        4.233
```

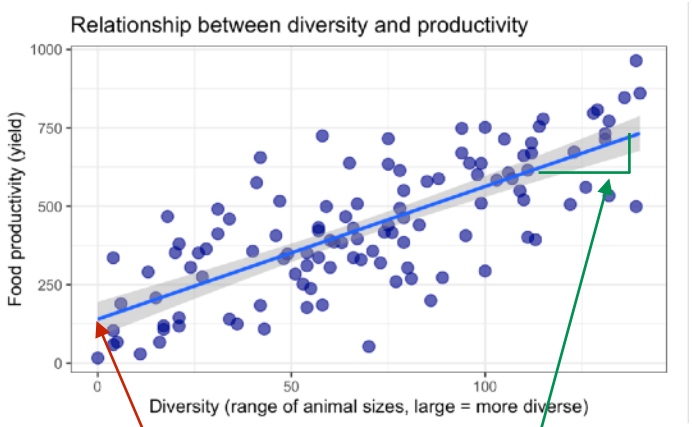
```
> model2 <- lm(yield ~ diversity + area, data=d)
(Intercept)    diversity          area
   -73.1733       1.3900        0.9872
```

Let's run the model

Notice that adding another variable changes everything!

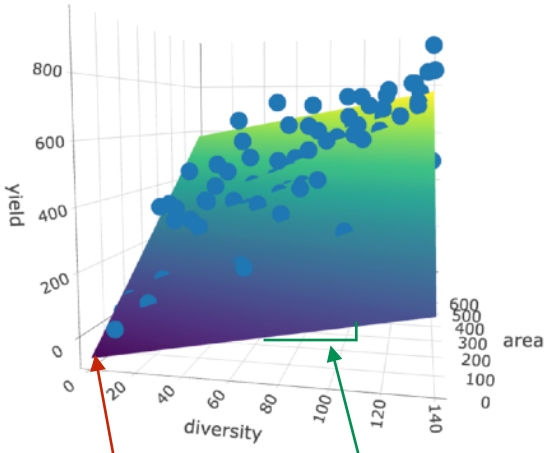
This is for a similar reason that adding variables to an ANOVA changes things: additional variables explain some of the variance. In a multiple regression with two variables, you're trying to find the plane that best fits *both of them*. The interpretation of the coefficients changes accordingly.

1 variable version (from Day 1)




```
> model1 <- lm(yield ~ diversity, data=d )
(Intercept)    diversity
    139.877        4.233
```

2 variable version (just now)




```
> model2 <- lm(yield ~ diversity + area, data=d)
(Intercept)    diversity          area
   -73.1733       1.3900        0.9872
```

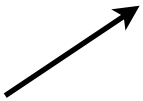
Can even extend to interaction terms

Yield $= b_1$* Diversity $+ b_2$* Area $+ b_3$* Diversity*Area $+ b_0$

$$Y_i = b_1 X_{1i} + b_2 X_{2i} + b_3 X_{1i} X_{2i} + b_0 + \epsilon_i$$

The regression model, using variables instead of words

```
yield ~ diversity + area + diversity:area
```

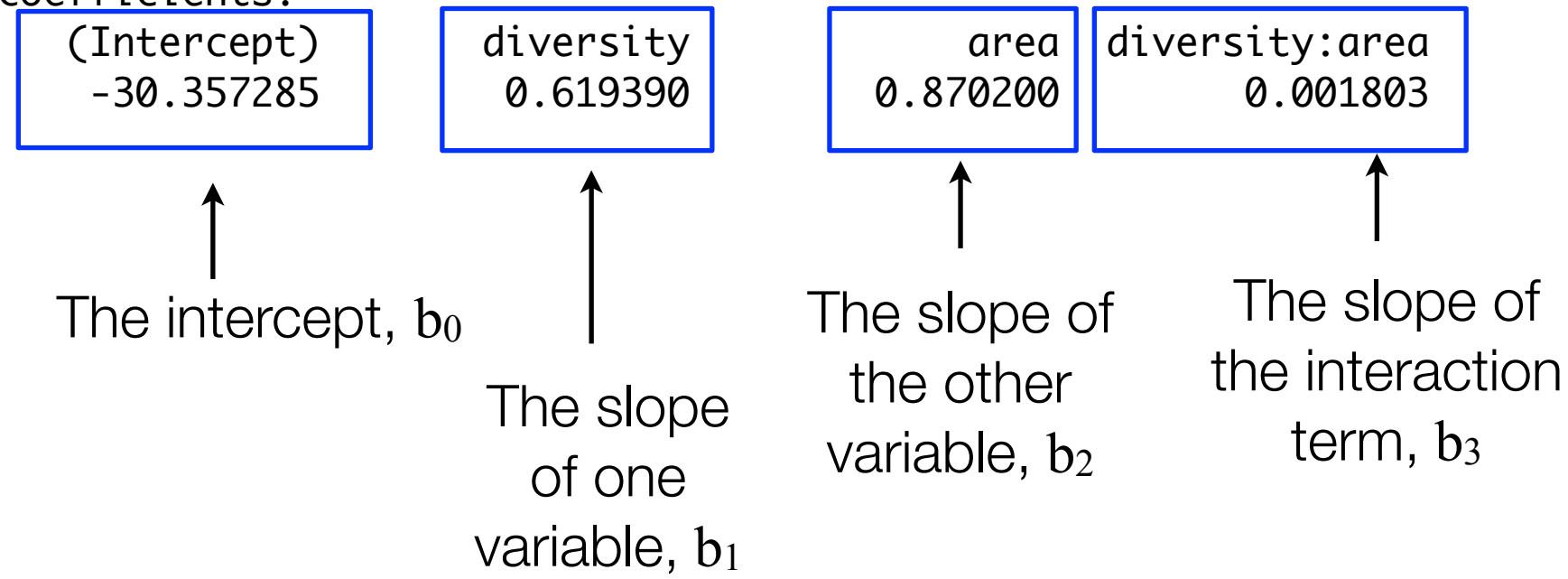


The R formula that describes this model

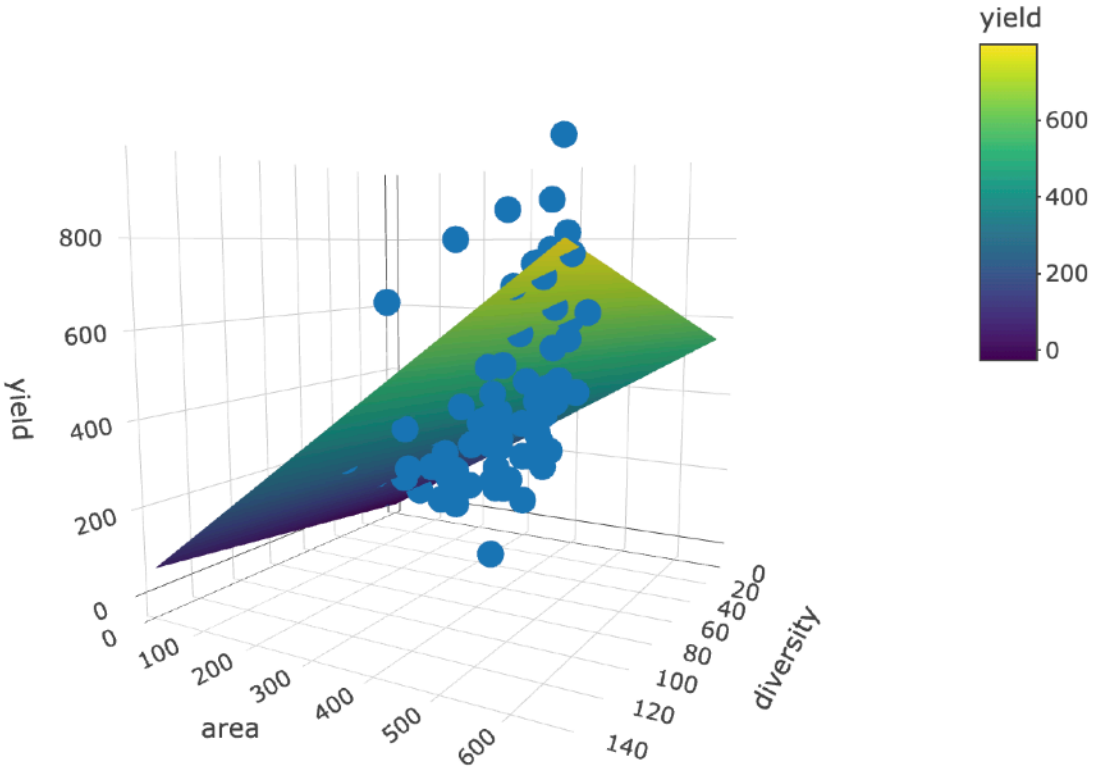
Can even extend to interaction terms

```
> model3 <- lm(yield ~ diversity + area + diversity:area, data=d)
> model3
Call:
lm(formula = yield ~ diversity + area + diversity:area, data = d)

Coefficients:
  (Intercept)     diversity          area  diversity:area
    -30.357285      0.619390      0.870200      0.001803
```



Visualising it



week9day2plots.html

(This is the second plot on there, showing the interaction)
The coefficient is tiny so it doesn't look that different!

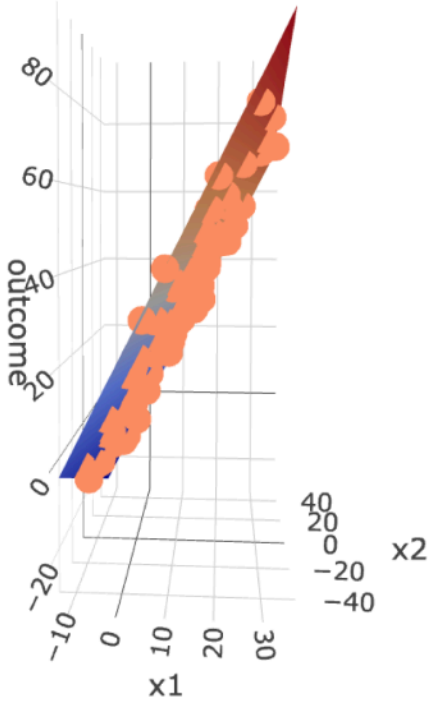
That had hardly any interaction… so what does a real interaction look like?

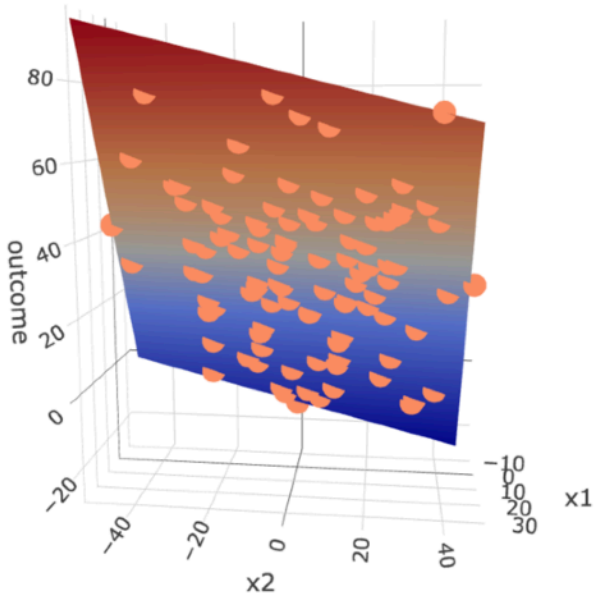
Let's generate our own data, cleanly, so we can see

Step 1. Simple multiple regression with no residuals and no interaction

```
x1 <- rnorm(n=100,mean=10,sd=10)
x2 <- rnorm(n=100,mean=0,sd=20)
outcome <- 10 + 2*x1 - 0.25*x2
```

$$Y = 2X_1 - 0.25X_2 + 10$$





week9day2plots.html

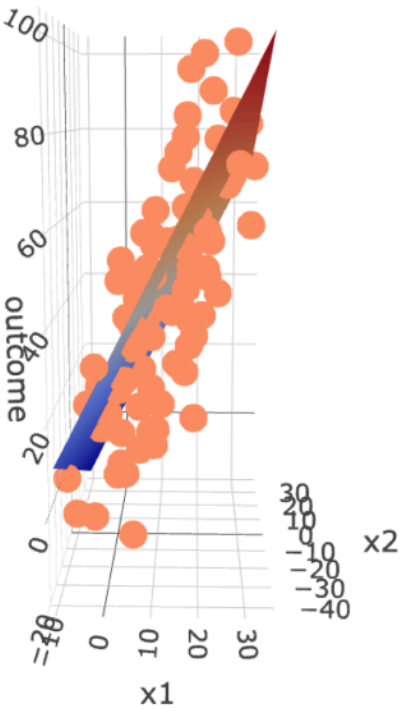
You can see that it is a *tilted flat* plane in both the X_1 and X_2 dimensions

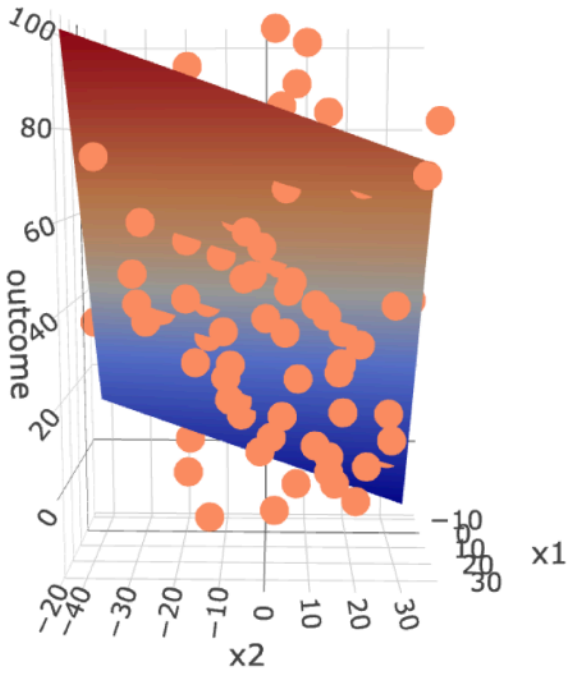
Let's generate our own data, cleanly, so we can see

Step 2. Same multiple regression with residuals and no interaction

```
x1 <- rnorm(n=100,mean=10,sd=10)
x2 <- rnorm(n=100,mean=0,sd=20)
outcome <- 10 + 2*x1 - 0.25*x2 + rnorm(n=100,mean=5,sd=20)
```

$$Y = 2X_1 - 0.25X_2 + 10 + \epsilon$$





week9day2plots.html

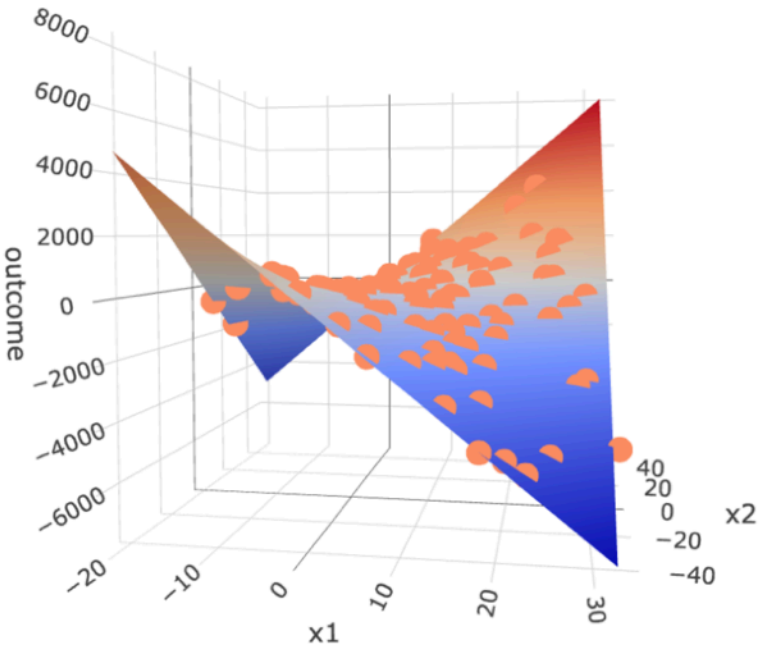
You can see that it is a *tilted flat* plane in both the X_1 and X_2 dimensions, and the dots don't fit on the plane

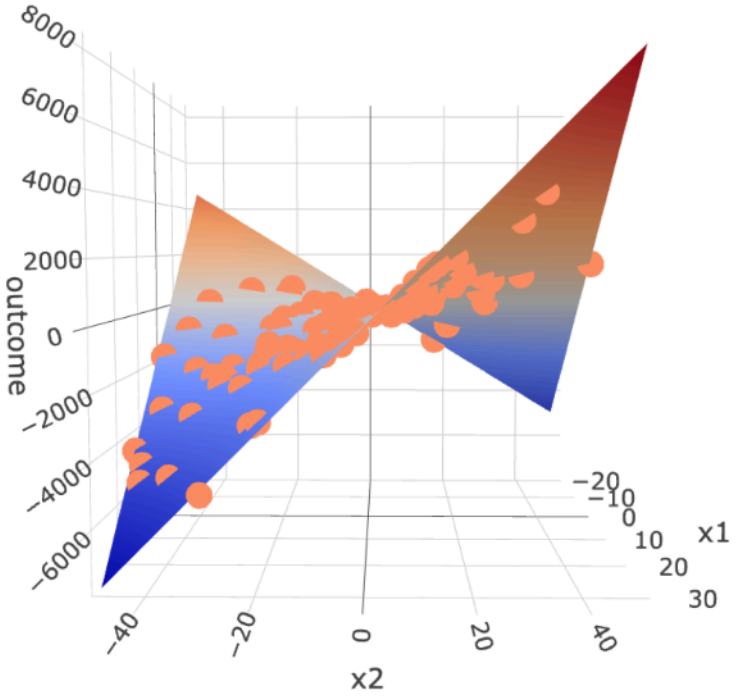
Let's generate our own data, cleanly, so we can see

Step 3. Multiple regression with same predictors, no residuals and interaction

```
x1 <- rnorm(n=100,mean=10,sd=10)
x2 <- rnorm(n=100,mean=0,sd=20)
outcome <- 10 + 2*x1 - 0.25*x2 + 5*x1*x2
```

$$Y = 2X_1 - 0.25X_2 + 5X_1X_2 + 10$$





week9day2plots.html

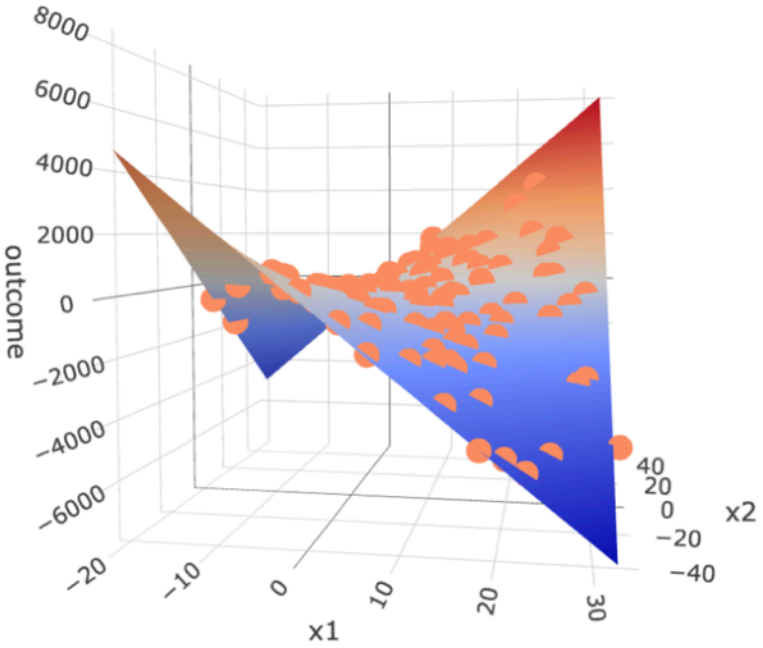
You can see that it is a *curved* plane in both the X_1 and X_2 dimensions - like a folded piece of paper

Let's generate our own data, cleanly, so we can see

Step 3. Multiple regression with same predictors, no residuals and interaction

```
x1 <- rnorm(n=100,mean=10,sd=10)
x2 <- rnorm(n=100,mean=0,sd=20)
outcome <- 10 + 2*x1 - 0.25*x2 + 5*x1*x2
```

$$Y = 2X_1 - 0.25X_2 + 5X_1X_2 + 10$$



Conceptually, an interaction describes the situation where the relationship between one predictor and the outcome depends on the nature of the other predictor

(Put another way, you can't understand how X_1 is related to Y unless you also know what X_2 is, and vice versa)

week9day2plots.html

The sign of the interaction governs the relationship

Positive interaction term: when X_1 and X_2 are the same sign (i.e., both positive or both negative), Y is **high**. It is **low** if the signs are different.

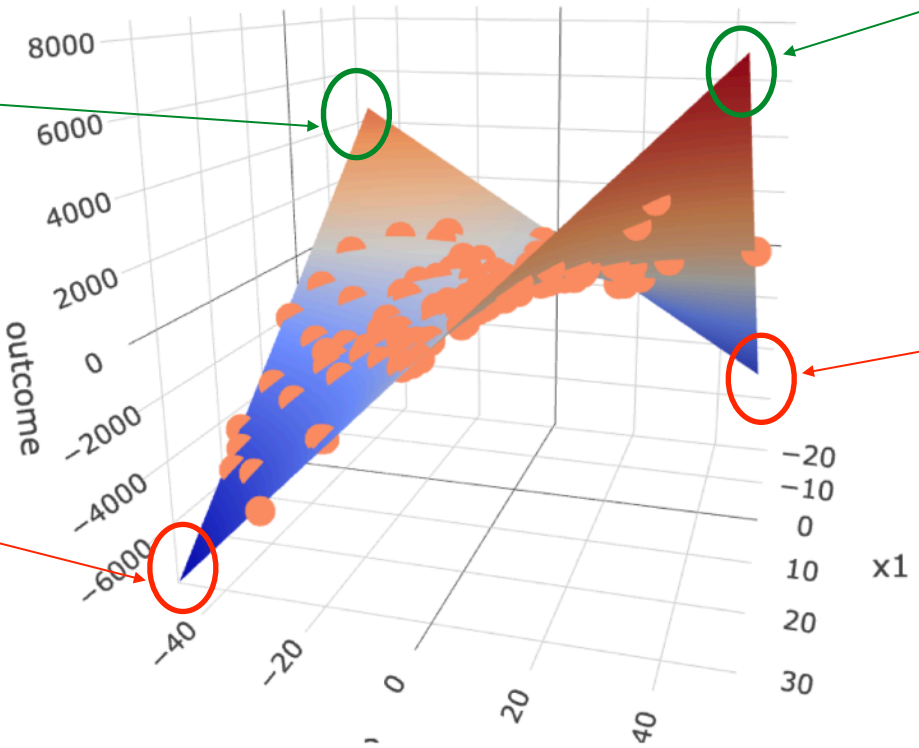
$$Y = 2X_1 - 0.25X_2 + 5X_1X_2 + 10$$

Both predictors are negative, Y is high:
$\{X_1 = -20, X_2 = -40, Y = 3980\}$

Both predictors are positive, Y is high:
$\{X_1 = 30, X_2 = 40, Y = 6060\}$

One predictor (X_2) is negative, Y is low:
$\{X_1 = 30, X_2 = -40, Y = -5920\}$

One predictor (X_1) is negative, Y is low:
$\{X_1 = -20, X_2 = 40, Y = -4040\}$



week9day2plots.html

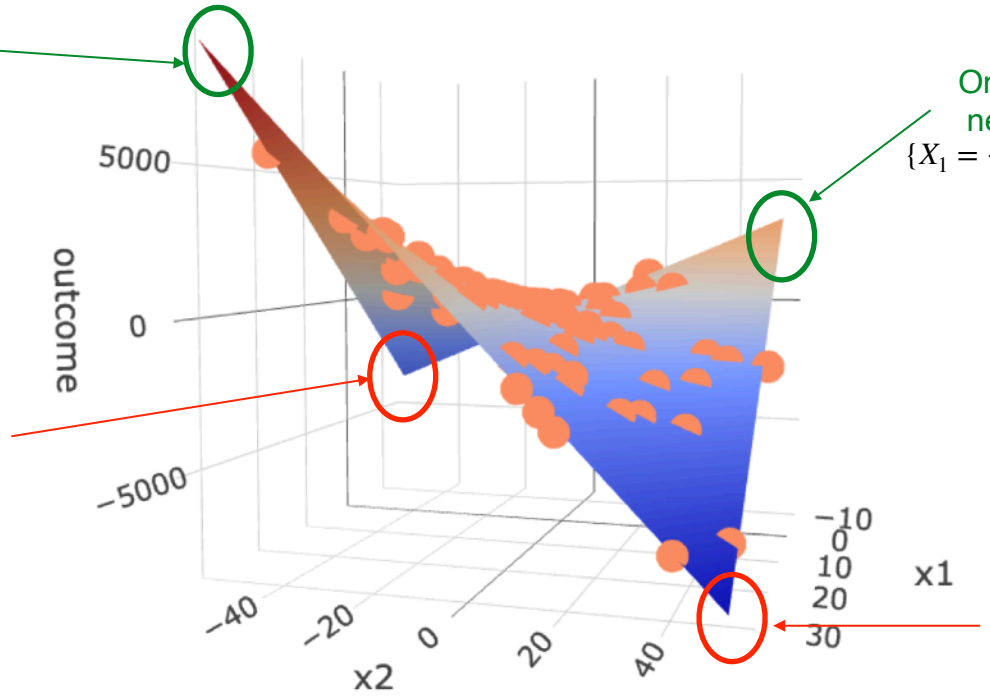
The sign of the interaction governs the relationship

Negative interaction term: when X_1 and X_2 are the same sign (i.e., both positive or both negative), Y is **low**. It is **high** if the signs are different.

$$Y = 2X_1 - 0.25X_2 - 5X_1X_2 + 10$$

One predictor (X_2) is negative, Y is high:
$\{X_1 = 30, X_2 = -40, Y = 6080\}$

One predictor (X_1) is negative, Y is high:
$\{X_1 = -10, X_2 = 40, Y = 1980\}$

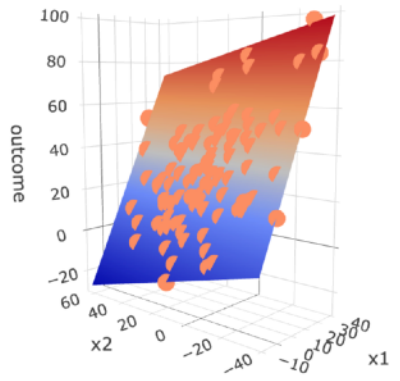
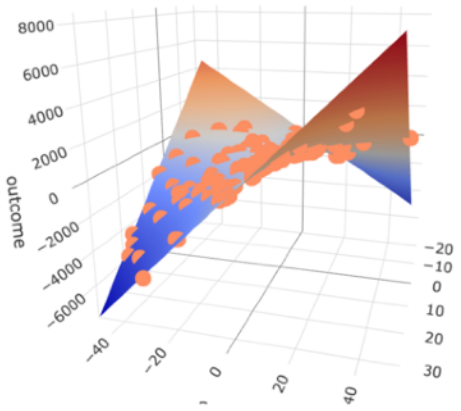
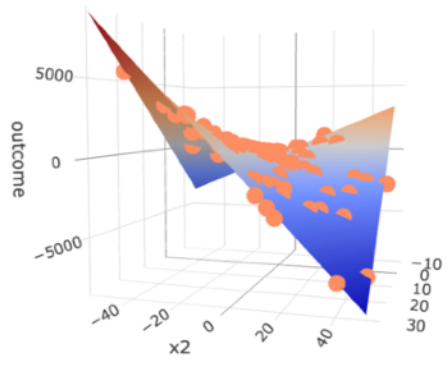


Both predictors are negative, Y is low:
$\{X_1 = -10, X_2 = -40, Y = -2000\}$

Both predictors are positive, Y is low:
$\{X_1 = 30, X_2 = 40, Y = -5940\}$

week9day2plots.html

Summary

No interaction	Positive interaction	Negative interaction
Tilted plane	Curved plane	Curved plane
Y depends on predictors independently	Y is high when predictors have the same sign	Y is low when predictors have the same sign
		

Exercises are in w9day2exercises.Rmd