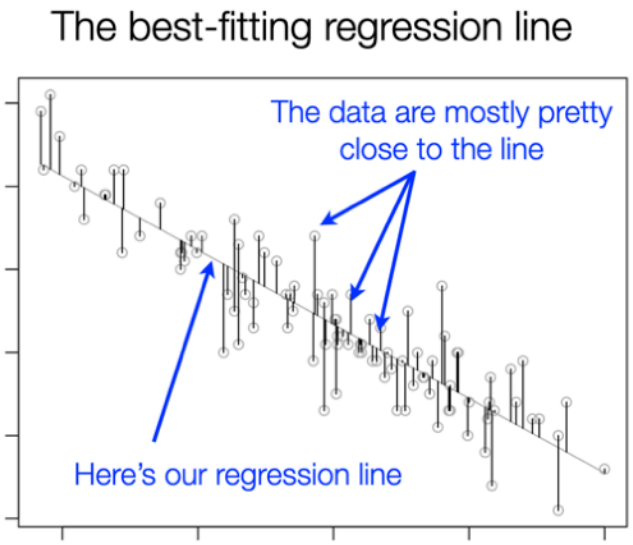
Comparing two numeric variables: More about linear regression

Research Methods for Human Inquiry

Andrew Perfors

Last time

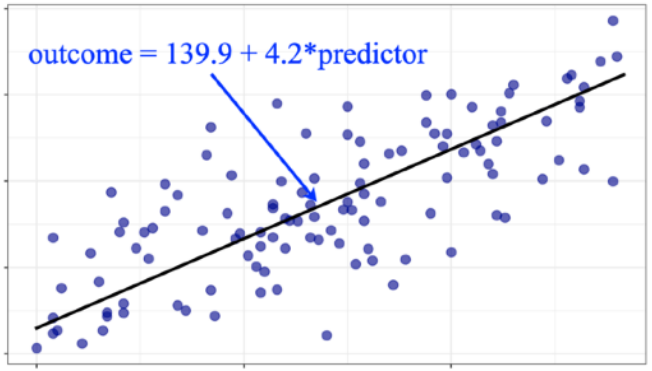
Regression finds the line of best fit by minimising the distance between the data points and the line



$$SS_{res} = \sum_i (Y_i - \hat{Y}_i)^2$$

This distance is the residual sum of squares

```
model <- lm(outcome ~ predictor, data=d )

Coefficients:
(Intercept)    predictor
    139.877        4.233
```



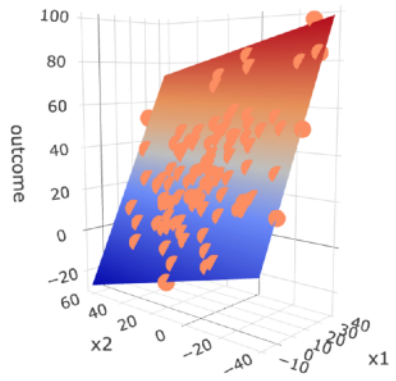
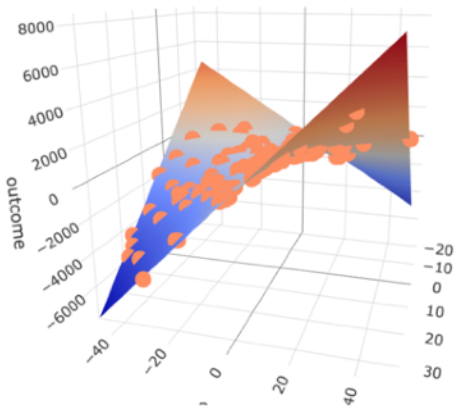
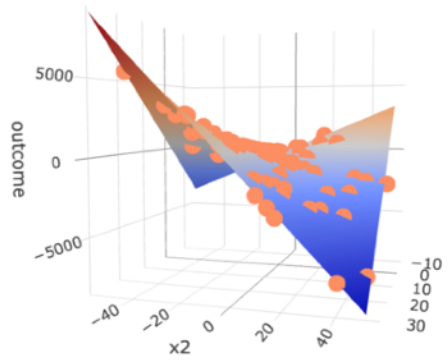
Can have multiple variables, corresponding to a multi-dimensional regression line, as well as interactions

```
M <- lm(outcome ~ pred1 + pred2, data=d )
```

$$\text{Outcome} = b_2 * \text{pred1} + b_1 * \text{pred2} + b_0$$

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

Last time

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
		

We learned it's not just diversity, but area that matters too. But which matters more? Are they both significant? If it's just that larger land has more diversity but it's the land size that's driving it, that's not useful.

Which variables matter the most? How do we quantify this?

How do we test significance?

Now: let's construct a test

1) A diagnostic test statistic, *T*
   F statistic

2) Sampling distribution of *T* if the null is true
   F distribution

3) The observed *T* in your data

4) A rule that maps every value of *T* onto a decision (accept or reject H0)

These are exactly analogous to the ANOVA

Test #1: Is the overall regression equation “significant”?

We want a hypothesis test to check whether the performance of the *overall* full model (with a distinct slope for each predictor variable) is better than what you'd expect if the null hypothesis were true.

$$H_0 : Y_i = b_0 + \epsilon_i$$

- Null hypothesis:
  - No relationship between the predictors and outcome
  - Equivalently, all slope parameters are zero

$$H_1 : Y_i = b_1 X_{1i} + \ldots + b_k X_{ki} + b_0 + \epsilon_i$$

- Alternative hypothesis:
  - The relationship between the predictors and outcome matches the regression model

Test #1: Is the overall regression equation “significant”?

We can test this using the same F-test that we used in our ANOVA. For a model with K predictor variables we have two sources of variance

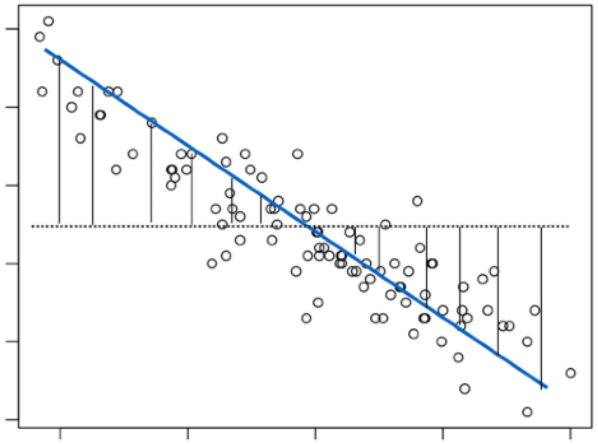
- **Model sum of squares (SS_m):**

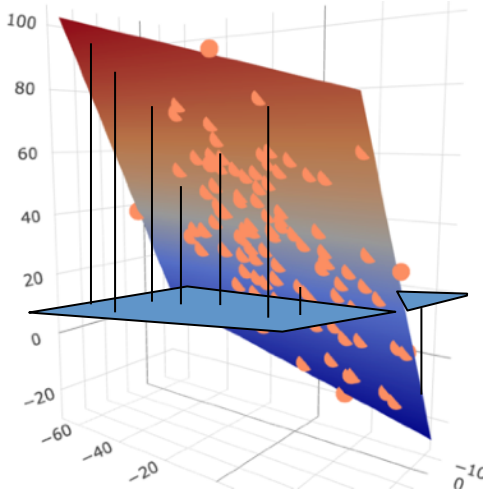
- **Residual sum of squares (SS_r):**

Test #1: Is the overall regression equation "significant"?

Model sum of squares (SS_m): the difference between regression line predictions and the mean Y value (df = the number of predictors K)

- Analogous to between-groups variance



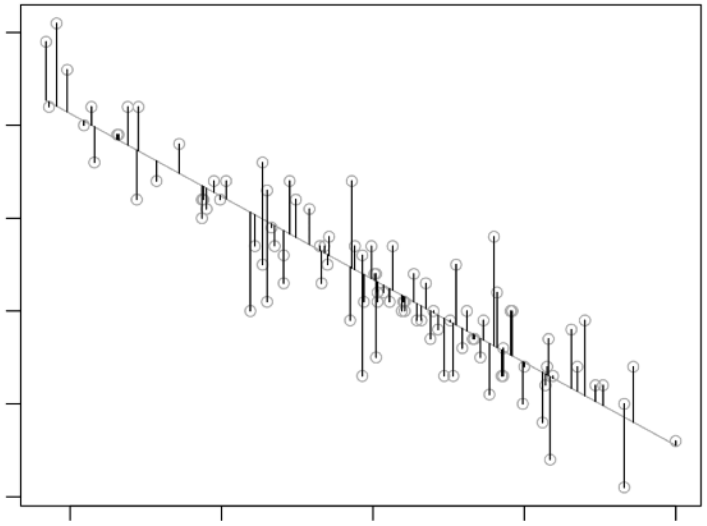


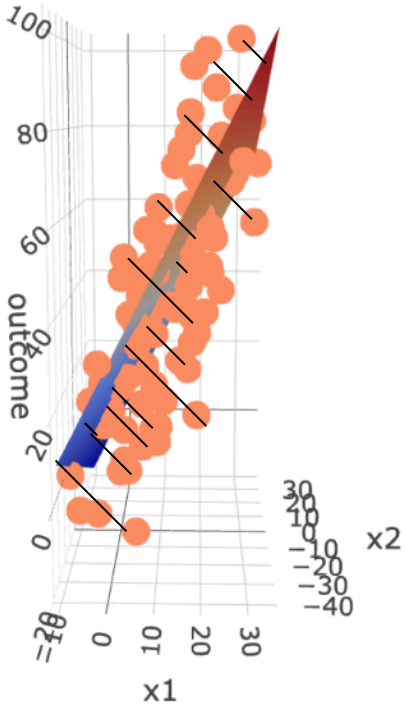
The steeper the slope (i.e., the stronger the relationship between predictor(s) and outcome), the larger SS_m *is*

Test #1: Is the overall regression equation “significant”?

Residual sum of squares (SS_r): the difference between the data and the regression line predictions (df = N - K - 1)

- Analogous to within-groups variance





The farther away the datapoints are from the prediction (i.e., the more error there is), the larger SS_r is

Test #1: Is the overall regression equation "significant"?

We can test this using the same F-test that we used in our ANOVA. For a model with K predictor variables we have two sources of variance

- **Model sum of squares (SS_m):** the difference between regression line predictions and the mean Y value (df = number of predictors K)
  - Analogous to between-groups variance
- **Residual sum of squares (SS_r):** the difference between the data and the regression line predictions (df = $\mathrm{N} - \mathrm{K} - 1$)
  - Analogous to within-groups variance

Calculate the F-statistic as before.

Test #2: Is a specific predictor significant?

- Similar idea to before, but now we want to test the null hypothesis that a specific predictor k has no relationship to the outcome
  - Null hypothesis: $H_0 : b_k = 0$
  - Alternative hypothesis: $H_1 : b_k \neq 0$

- We can use t-tests for this (a different t-test for each predictor)
  - df $= N - K - 1$

Note that R does not do multiple corrections for the t-tests in a regression. That's much less of a problem here because (a) they are often more well-motivated theoretically (and hence more like planned comparisons) and (b) they don't explode in number, because it doesn't do all the pairwise tests, just looks at each variable compared to the outcome

Effect size measure

- How well does the overall regression model actually account for the outcome variable?
  - i.e., how closely does the model predictions $\hat{Y}$ match up to the actually observed Y values?

- Intuitively, it should depend on how much variance is accounted for by the model, compared to how much total variance there is

$$R^2 = 1 - \frac{SS_r}{SS_r + SS_m}$$

SS_r ← deviation between predictions and actual values

$SS_r + SS_m$ ← total variability in the outcome

Effect size measure

- How well does the overall regression model actually account for the outcome variable?
  - i.e., how closely does the model predictions $\hat{Y}$ match up to the actually observed Y values?

- Intuitively, it should depend on how much variance is accounted for by the model, compared to how much total variance there is

$$R^2 = 1 - \frac{SS_r}{SS_{tot}}$$

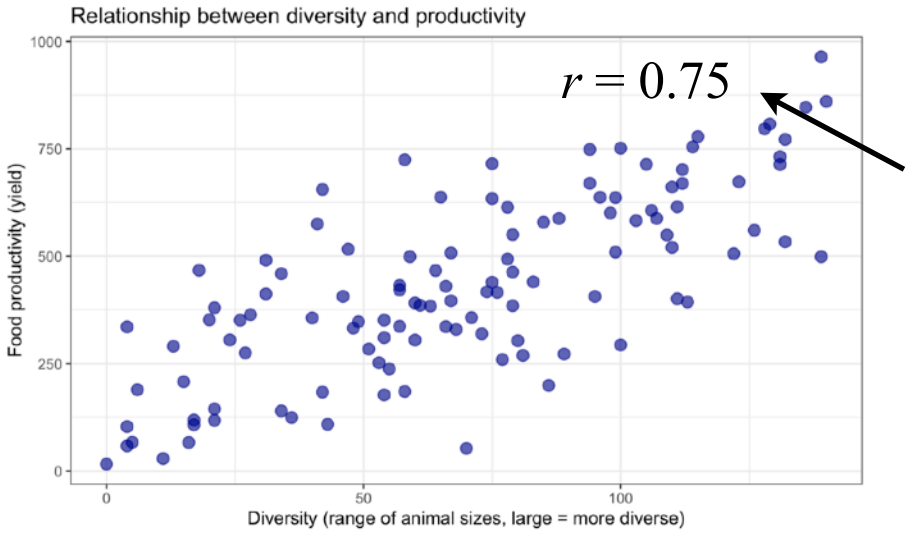
SS_r ← deviation between predictions and actual values

SS_{tot} ← total variability in the outcome

R^2 is one when there is no deviation between the predictions and actual values
R^2 is zero when the deviation is the same as the total variability

Effect size measure

- Like η^2, this can be interpreted as the proportion of the variance in the outcome accounted for by the model
  - So if a model with a single predictor X has an R^2 of 0.52 means that 52% of the variance is due to X
- This is also *the same thing* as the Pearson correlation r, squared (which we talked about in the first video)



suggests we should find an R^2 of $(0.75)^2 = 0.56$

incidentally, this also means that the significance of that predictor in the regression is also the significance of the correlation with that predictor!

Doing a regression in R

It’s just like ANOVA...

Assume you've already run your regression using the `lm()` function, and stored the results as a variable called (say) `model2`

All you need is a `summary()` of this variable

Here's an example...

Doing a regression in R

```
> summary(model2)

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

Residuals:
    Min      1Q  Median      3Q     Max
-365.78  -73.57   -9.88   74.91  297.04

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -73.1733    36.4721  -2.006  0.04724 *
diversity     1.3900     0.4765   2.917  0.00427 **
area          0.9872     0.1314   7.511 1.53e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 116.2 on 112 degrees of freedom
Multiple R-squared:  0.707,   Adjusted R-squared:  0.7018
F-statistic: 135.1 on 2 and 112 DF,  p-value: < 2.2e-16
```

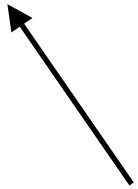
Doing a regression in R

```
> summary(model2)

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

This part is R reminding us what variables are involved in the regression, and what data frame they're stored in

Doing a regression in R

```
> summary(model2)


Residuals:
    Min      1Q  Median      3Q     Max
-365.78  -73.57   -9.88   74.91  297.04
```



blah blah don’t worry about these

Doing a regression in R

This is a summary of all of the regression coefficients, and the associated t-tests (Test #2 earlier) to check if they're significantly different from 0

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -73.1733    36.4721  -2.006  0.04724 *
diversity     1.3900     0.4765   2.917  0.00427 **
area          0.9872     0.1314   7.511 1.53e-11 ***
```

The estimates of the slope and intercept, i.e., the b values

Standard errors around b

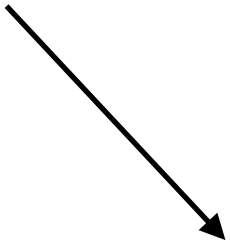
t-statistics for each predictor

p-value for each predictor

$$\hat{Y} = 1.39X_1 + 0.98X_2 - 73.2$$

Doing a regression in R

blah blah the only thing that matters
here is to note the degrees of freedom
for our t-tests on specific variables



```
Residual standard error: 116.2 on 112 degrees of freedom
```

Doing a regression in R

Remember R^2 is our measure of effect size (% of variance accounted for by all of our predictor variables together, equivalent to η^2 in ANOVA)

This is an adjustment based on # of predictors, don't worry about it

```
Multiple R-squared:  0.707,   Adjusted R-squared:  0.7018
```

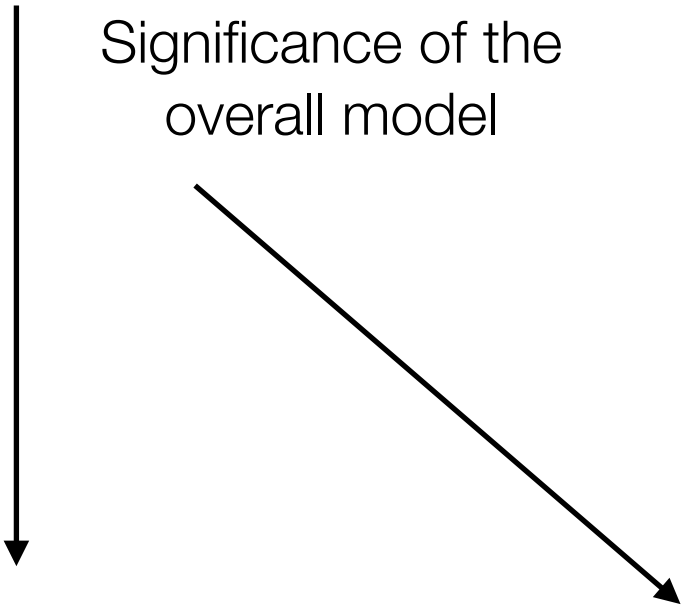
Doing a regression in R

Stat reference for the regression model:

$$F(2,112) = 135.1,\ p<.0001$$

Overall F statistic of the *overall* model (Test #1 earlier)

Significance of the overall model



```
F-statistic: 135.1 on 2 and 112 DF,  p-value: < 2.2e-16
```

How *do* I report the results?

- This is tricky…
  - There are different tests (t-tests, F-test) and different measures to think about (b-values, R^2 values), so you could do lots of different things.
- Advice on this varies
  - Some textbooks say “report everything, and have tables and stuff". That’s a LOT. It’s basically the entire R output.
- But that’s… kind of crazy. No journal in the world will let you include that much output, and nobody will read that
- Include what is important for understanding your findings. Means you have to think a bit

How *do* I report the results?

Example: this regression

We evaluated a linear model whose outcome variable was the yield of each plot of land and whose two predictors were the area of the land and the diversity of species farming the land (reflected in the range of sizes of those species). This model was statistically significant, $F(2,112) = 135.1$, $p<.0001$, accounting for 71% of the variance in yield. Area was a significant predictor, t(112)=-7.51, p<.0001, with a slope of 0.98, suggesting that each additional square metre resulted in an additional 0.98 units of yield. Diversity was also significant, t(112)=2.92, p=.004, with a slope of 1.39, suggesting that an increase in size differential between the species of 1cm resulted in an additional 1.39 units of yield.

How *do* I report the results?

Example: this regression

We evaluated a linear model whose outcome variable was the yield of each plot of land and whose two predictors were the area of the land and the diversity of species farming the land (reflected in the range of sizes of those species). This model was statistically significant, $F(2,112) = 135.1$, $p<.0001$, accounting for 71% of the variance in yield. Area was a significant predictor, $t(112)=-7.51$, $p<.0001$, with a slope of 0.98, suggesting that each additional square metre resulted in an additional 0.98 units of yield. Diversity was also significant, $t(112)=2.92$, $p=.004$, with a slope of 1.39, suggesting that an increase in size differential between the species of 1cm resulted in an additional 1.39 units of yield.

Describes the model (supplementary materials might also include the R command and version). Here I went into some detail about how the predictors were defined; sometimes you might do that before so you don't need to explain it so much here, but it definitely needs to be explained somewhere. Don't just use variable names without explanation, especially if they are unlikely to be meaningful to someone without the dataset.

How *do* I report the results?

Example: this regression

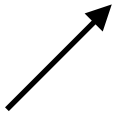
We evaluated a linear model whose outcome variable was the yield of each plot of land and whose two predictors were the area of the land and the diversity of species farming the land (reflected in the range of sizes of those species). This model was statistically significant, $F(2,112) = 135.1$, $p<.0001$, accounting for 71% of the variance in yield. Area was a significant predictor, $t(112)=-7.51$, $p<.0001$, with a slope of 0.98, suggesting that each additional square metre resulted in an additional 0.98 units of yield. Diversity was also significant, $t(112)=2.92$, $p=.004$, with a slope of 1.39, suggesting that an increase in size differential between the species of 1cm resulted in an additional 1.39 units of yield.

Is the model as a whole significant? How much variance does it account for?

How *do* I report the results?

Example: this regression

We evaluated a linear model whose outcome variable was the yield of each plot of land and whose two predictors were the area of the land and the diversity of species farming the land (reflected in the range of sizes of those species). This model was statistically significant, $F(2,112) = 135.1$, $p<.0001$, accounting for 71% of the variance in yield. Area was a significant predictor, t(112)=-7.51, p<.0001, with a slope of 0.98, suggesting that each additional square metre resulted in an additional 0.98 units of yield. Diversity was also significant, t(112)=2.92, p=.004, with a slope of 1.39, suggesting that an increase in size differential between the species of 1cm resulted in an additional 1.39 units of yield.



Describe which predictors were significant and how to interpret them (you don't always need to include slopes but it's useful if you want to interpret what each predictor is doing). Be aware of units when you are doing your interpretation!

Standardised regression coefficients

- Sometimes your predictors are genuinely on the same scale (i.e., all sizes, etc)
- This makes it fairly easy to compare the corresponding regression coefficents:
  - If one is larger than another, we can conclude that it has a stronger relationship to the outcome variable
- This is not true in general.

In our study

```
yield ~ diversity + area

Blah blah

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -73.1733    36.4721  -2.006  0.04724 *
diversity     1.3900     0.4765   2.917  0.00427 **
area          0.9872     0.1314   7.511 1.53e-11 ***

blah blah blah
```

Both are significant, but which is more important?

Which is more important?

- Here's the regression equation...

$$\texttt{Yield} = (1.39 * \texttt{Diversity}) + (0.98 * \texttt{Area}) - 73.2$$

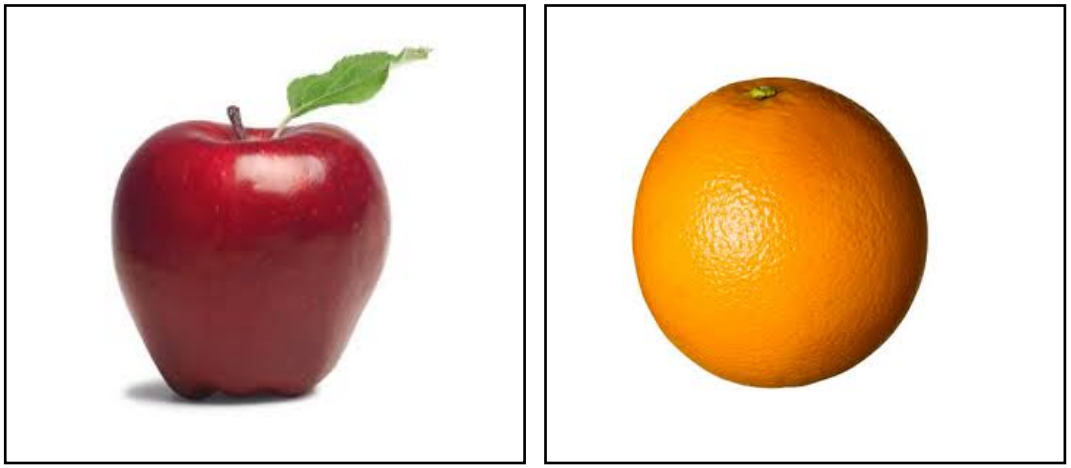
- It's tempting to think that Diversity has a stronger relationship to Yield than Area does
- This is misleading. Why?

Which is more important?

- These regression coefficients only tell you about the overall relationship between the raw scores.
- Our variables are on different scales:
  - `area` has mean 417 and standard deviation 136
  - `diversity` has mean 70 and standard deviation 38
- We cannot directly compare raw regression coefficients when the variables are so different to one another

We've seen this problem before

- We want to compare apples to oranges.



- Last time, the solution was to standardise all our variables (i.e., convert to z-scores)
- Can we repeat this trick?

Standardised coefficients

- Suppose we converted all our variables to standard scores before running our regression?
  - All of our variables would be on the same scale
  - So we could directly compare the resulting regression coefficients to one another

- These "**standardised coefficients**" are usually denoted β (beta) instead of b

Standardised coefficients

- The `standardCoefs()` function [`lsr` package] will extract standardised regression weights for you
- The only input you need to specify is an `lm` object.

```
> library(lsr)
> standardCoefs(model2)
                   b      beta
diversity 1.3900178 0.2456155
area      0.9871627 0.6324024
```

These are our standardised coefficients... as it turns out, area makes more difference than diversity (which makes a lot of sense!)

Unstandardised coefficients are useful for interpreting the slopes of the regression line (what a change in one variable means for the outcome). Standardised coefficients are useful for comparing them to each other

What can we conclude, then?

What can we conclude, then?

If people from Bunnyland and Otherland can farm together, we might be able to use the land appropriately and no longer be hungry!

What can we conclude, then?

Exercises are in w9day2exercises.Rmd