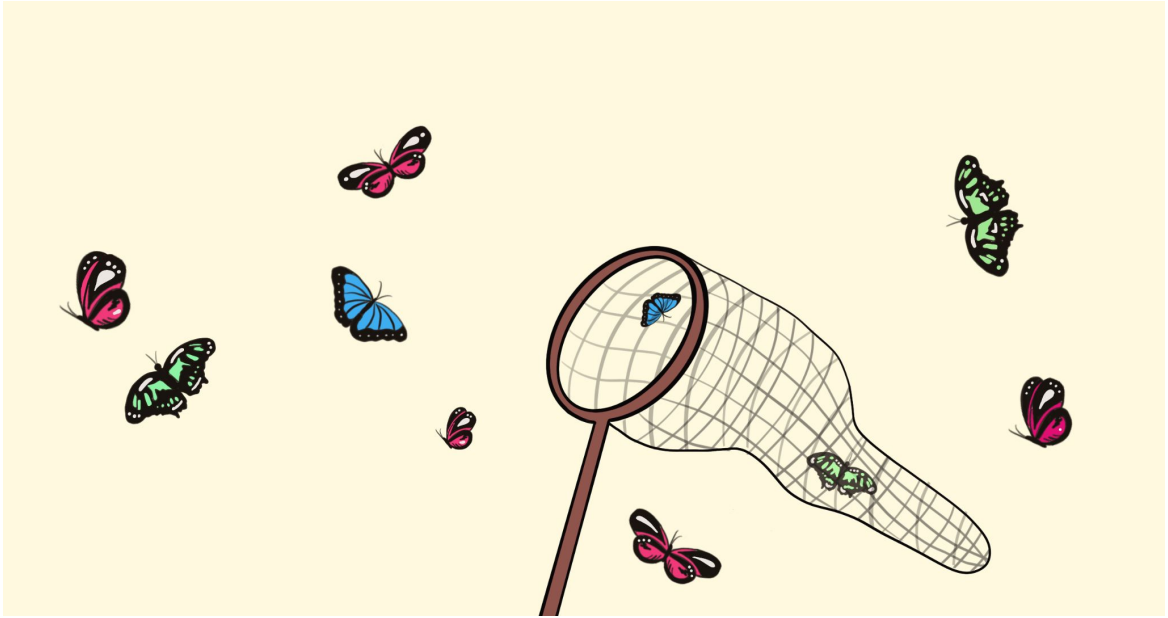
Week 5 Lecture 2:
Linear regression inference

EDS 222: *Statistics for Environmental Data Science*



Ocean acidification

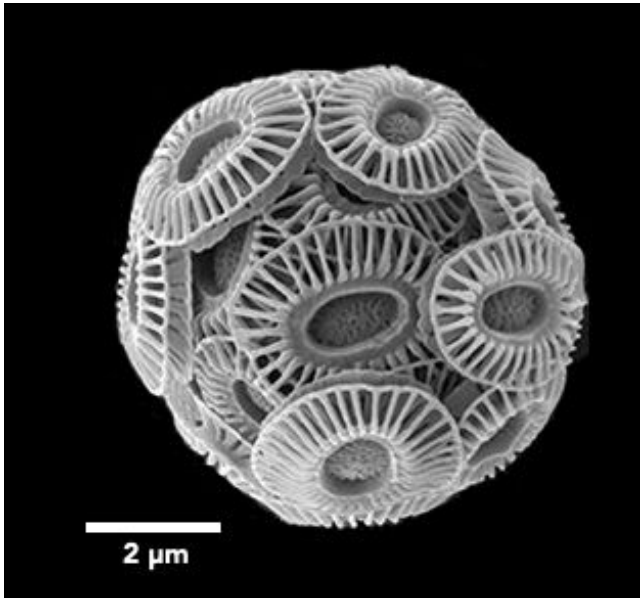
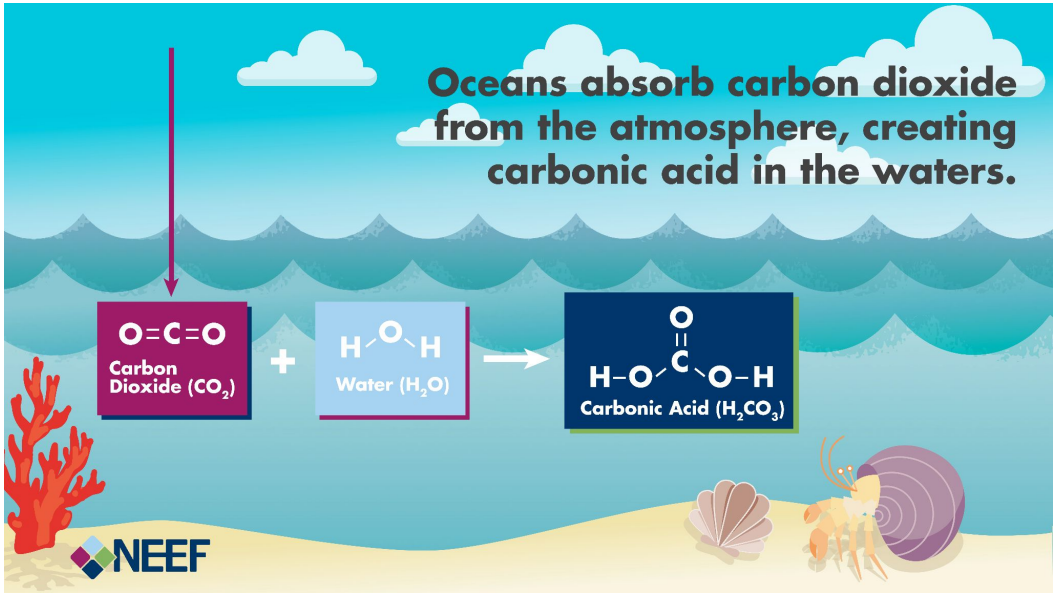
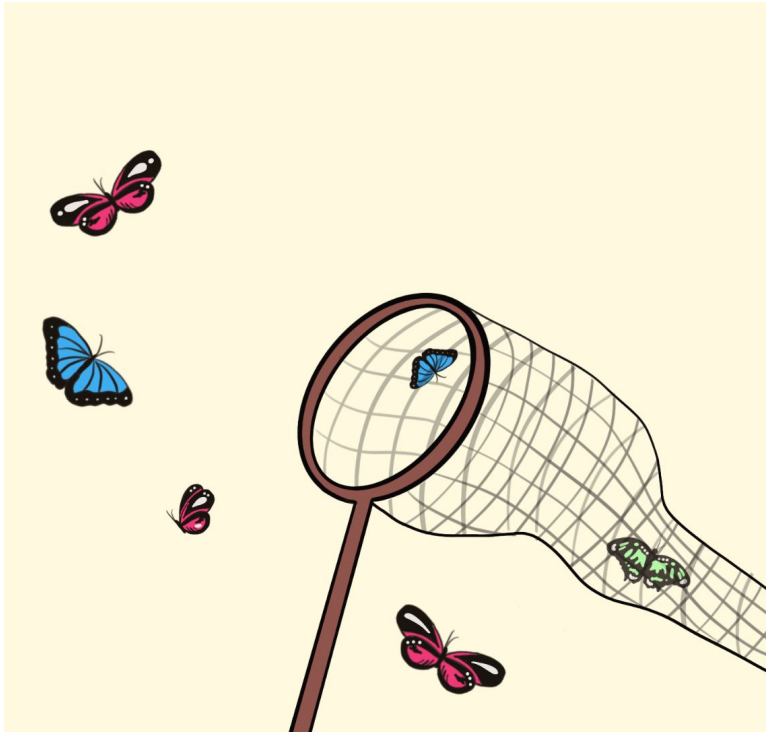


Photo: Jeremy Young



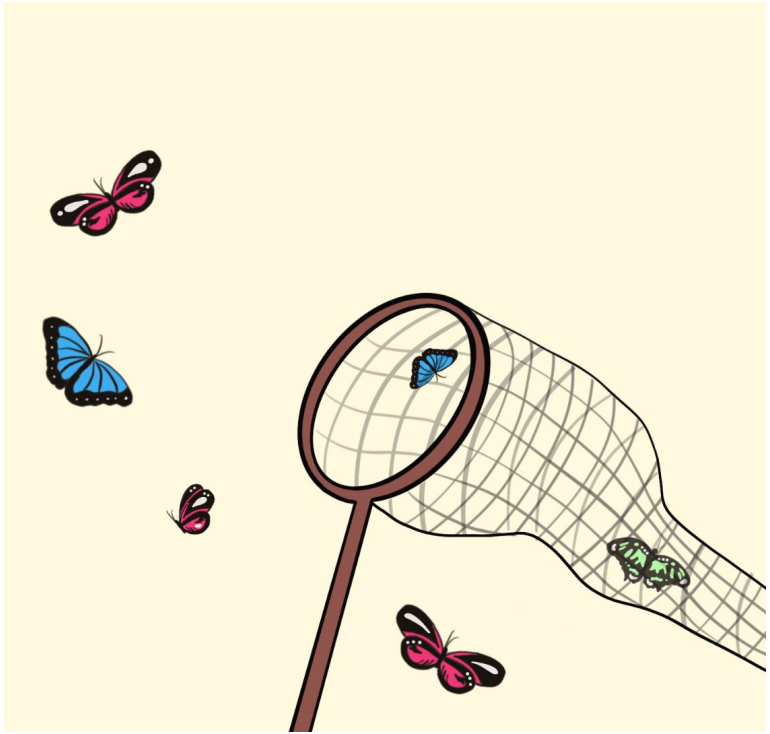
Today’s agenda

- ➔ Variation in regression
- ➔ Hypothesis testing
- ➔ Confidence intervals

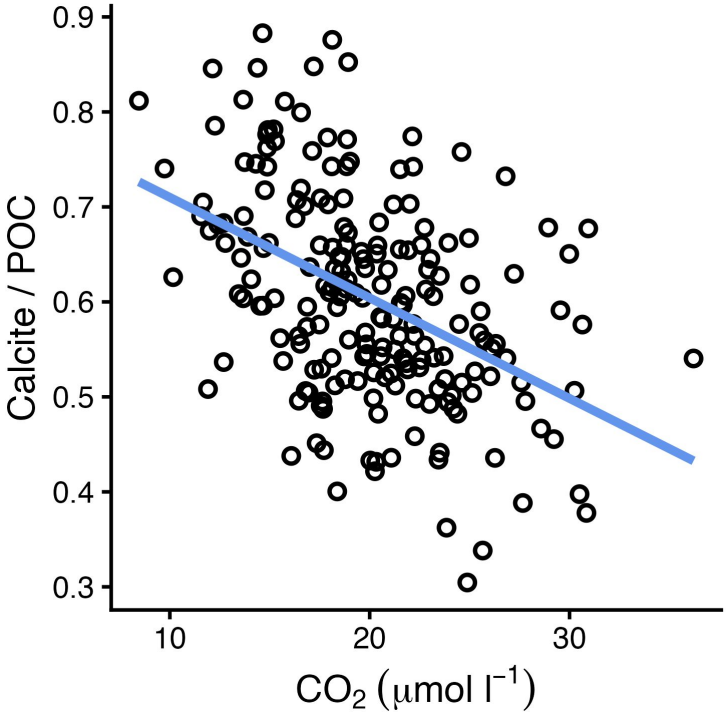


Today’s agenda

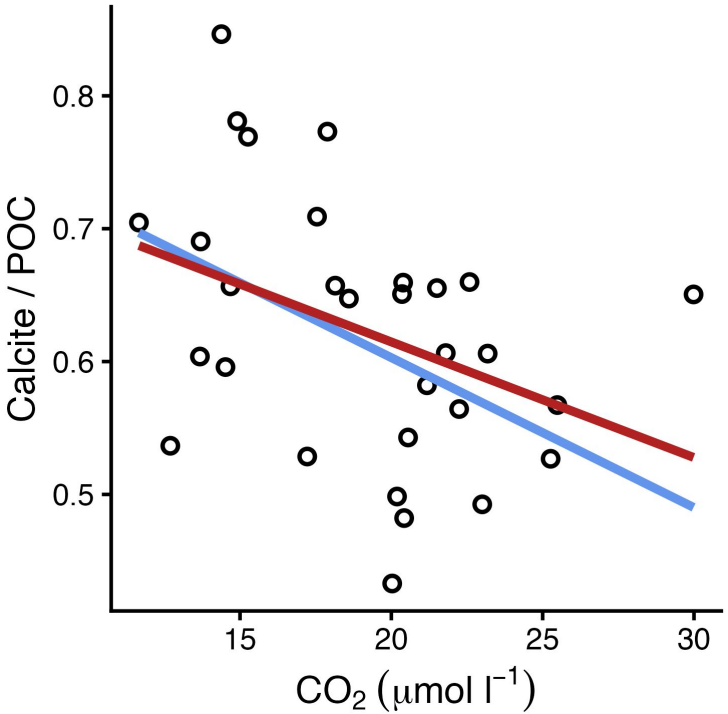
- ➔ **Variation in regression**
- ➔ Hypothesis testing
- ➔ Confidence intervals



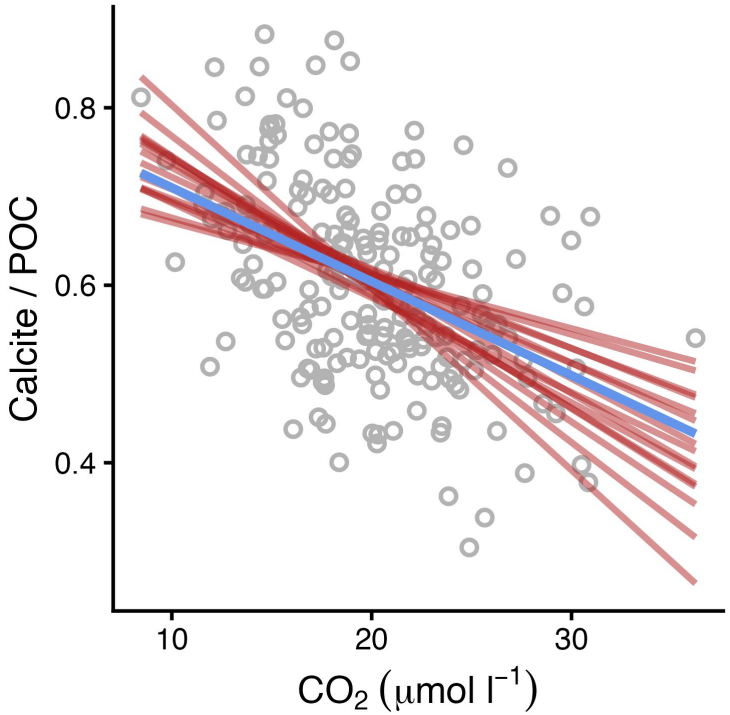
Population-level pattern



Draw a sample



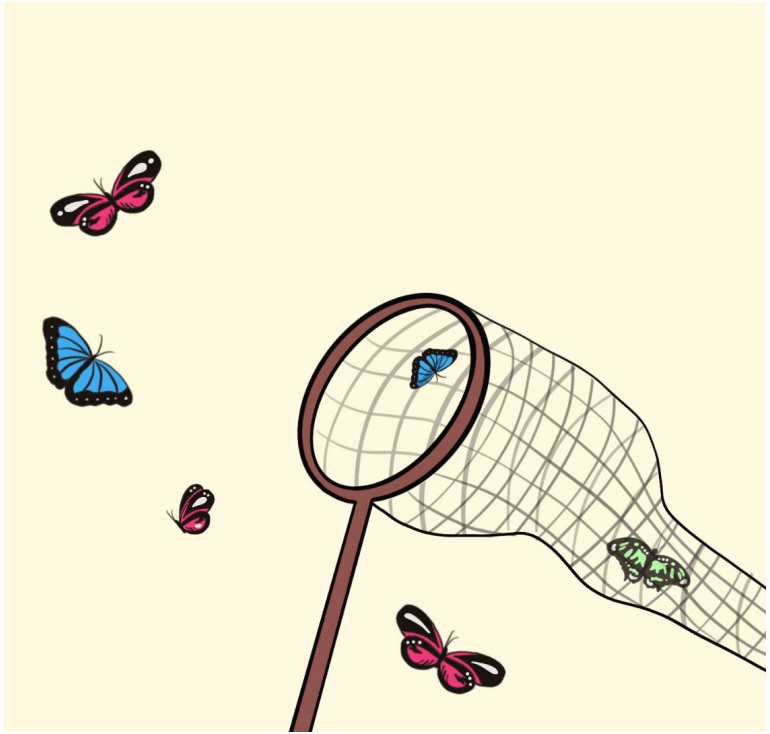
Draw a sample



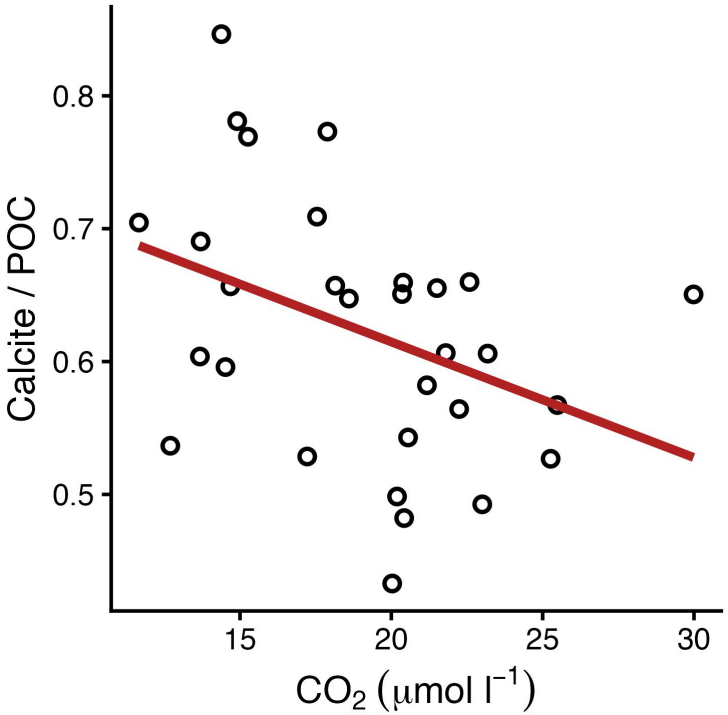
Variation in regression

Today’s agenda

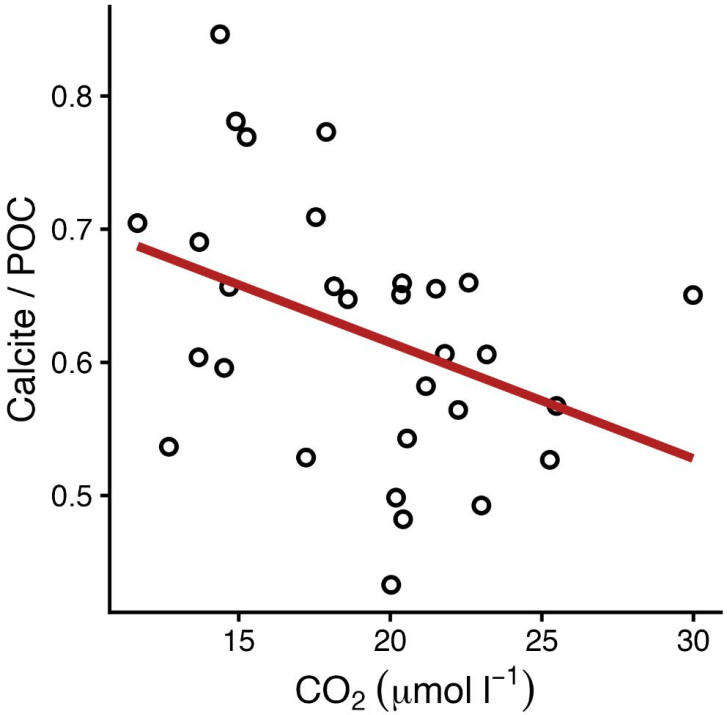
- ➔ Variation in regression
- ➔ **Hypothesis testing**
- ➔ Confidence intervals

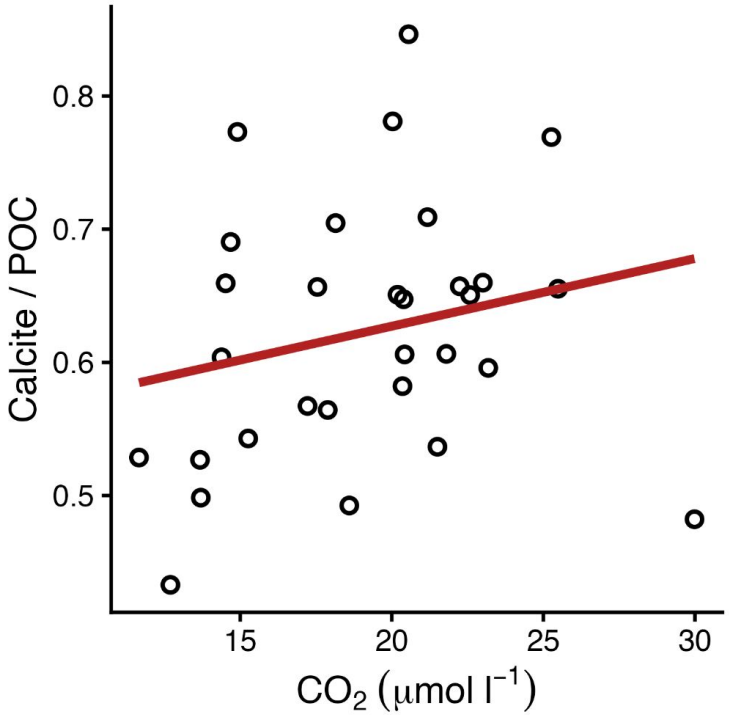


Hypothesis testing

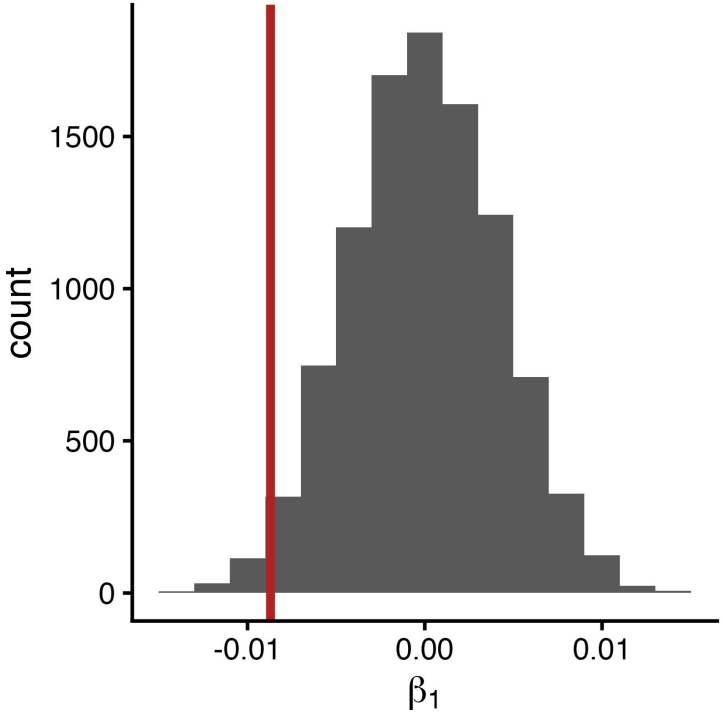


One permutation





Distribution of permutations



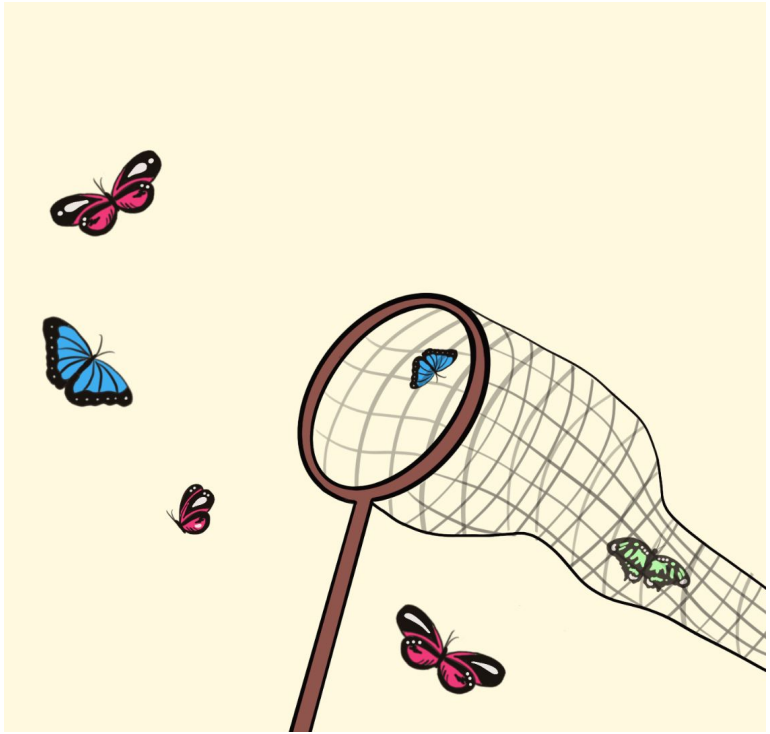
Mathematical model

```
Call:
lm(formula = calcite_poc ~ co2_umol_l, data = g_huxleyi_sample)

Residuals:
     Min       1Q   Median       3Q      Max
-0.18155 -0.06645  0.01227  0.05223  0.18277

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.788797   0.077446  10.185 6.41e-11 ***
co2_umol_l  -0.008702   0.003959  -2.198   0.0364 *
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.09167 on 28 degrees of freedom
Multiple R-squared:  0.1472,    Adjusted R-squared:  0.1167
F-statistic: 4.832 on 1 and 28 DF,  p-value: 0.03637
```

Hypothesis testing

Today’s agenda

- ➔ Variation in regression
- ➔ Hypothesis testing
- ➔ **Confidence intervals**

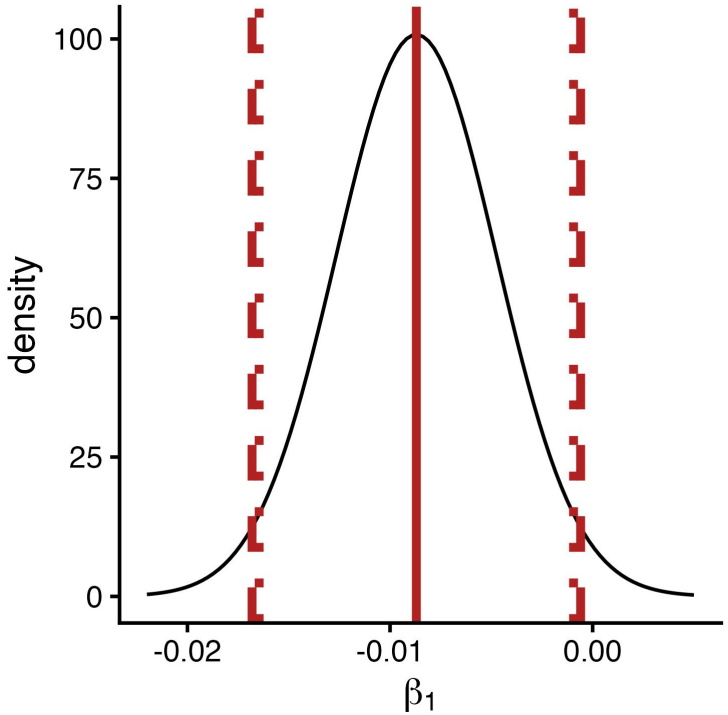


Confidence intervals

Coefficient CI

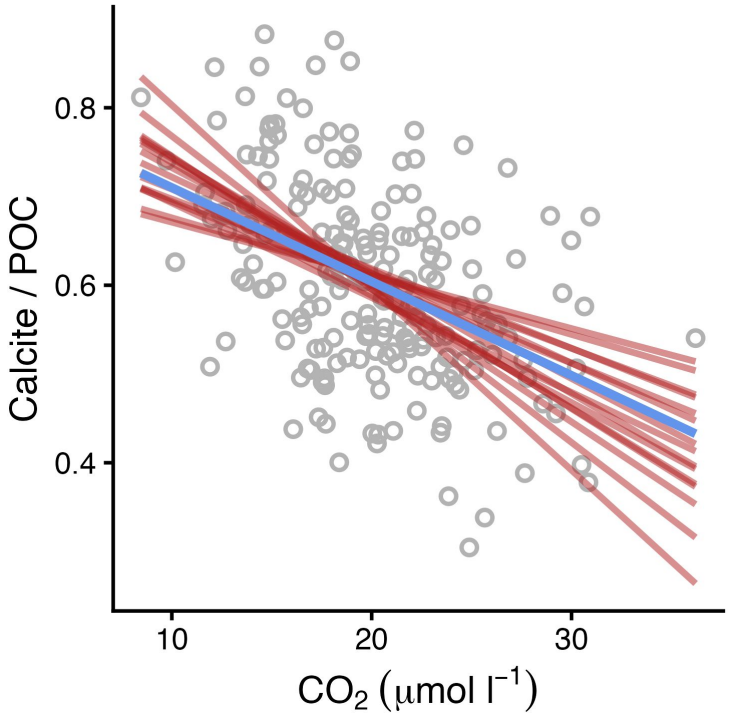
```
Call:
lm(formula = calcite_poc ~ co2_umol_l, data = g_huxleyi_sample)

Residuals:
     Min       1Q   Median       3Q      Max
-0.18155 -0.06645  0.01227  0.05223  0.18277

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.788797   0.077446  10.185 6.41e-11 ***
co2_umol_l  -0.008702   0.003959  -2.198   0.0364 *
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.09167 on 28 degrees of freedom
Multiple R-squared:  0.1472,    Adjusted R-squared:  0.1167
F-statistic: 4.832 on 1 and 28 DF,  p-value: 0.03637
```

Normal is pretty close

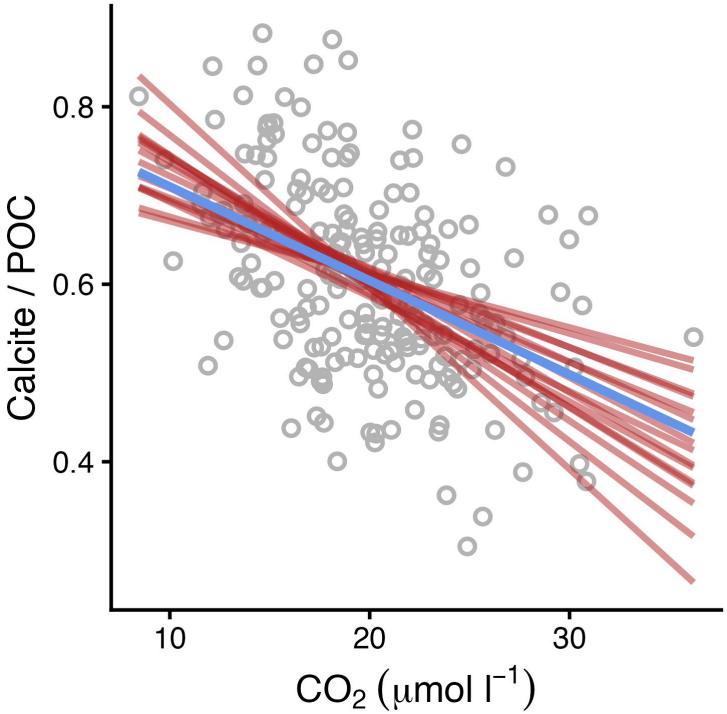


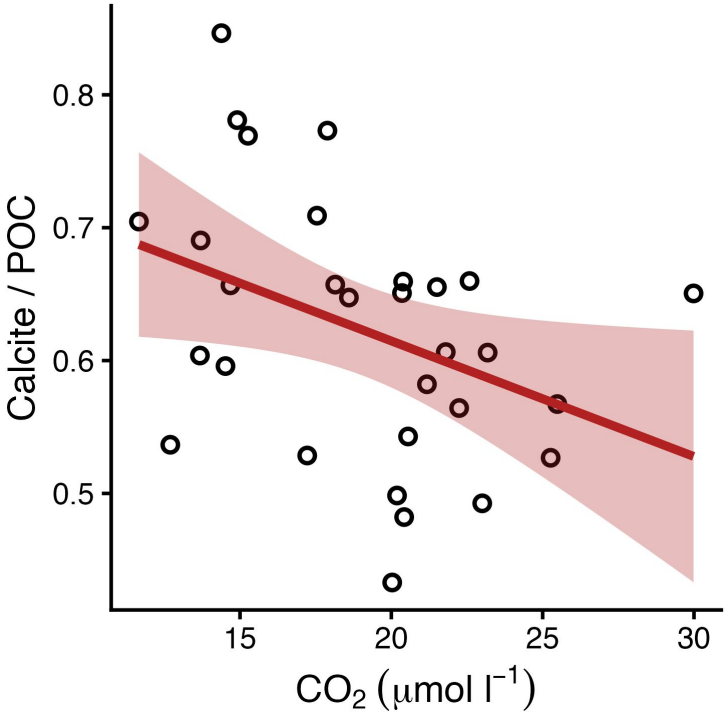
Coefficient CI

Mean response CI

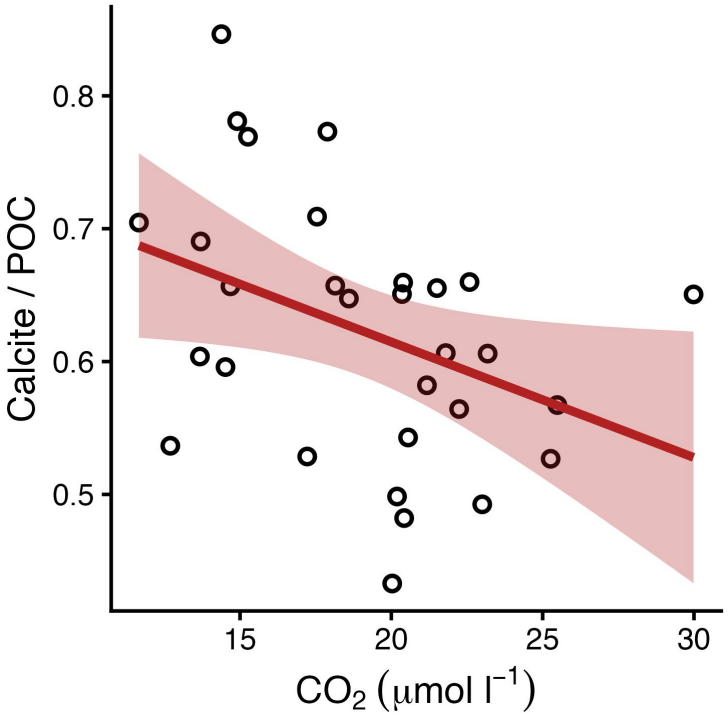


Mean response CI



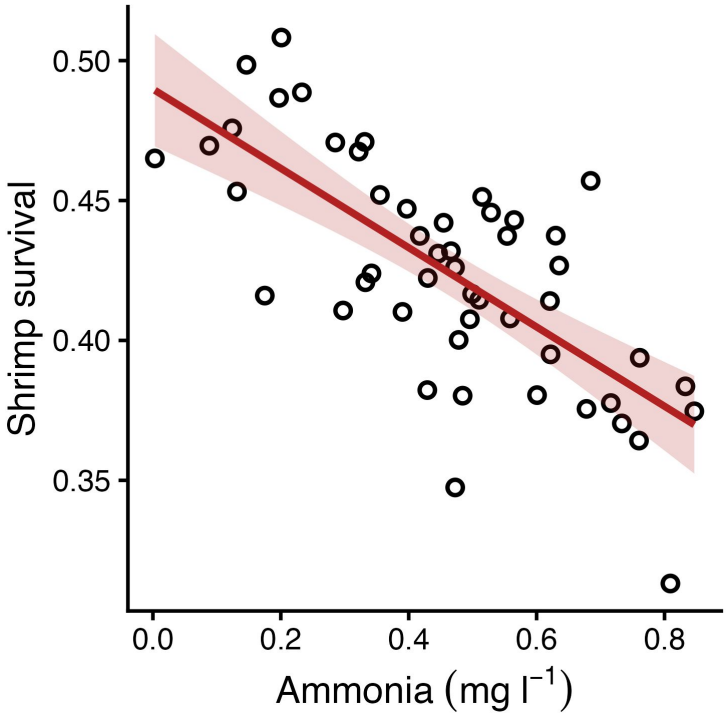


Mean response CI



Confidence intervals

Let's try it




```
> shrimp_aqua_lm <- lm(shrimp_survival ~ ammonia_mg_l, shrimp_aqua)
> summary(shrimp_aqua_lm)

Call:
lm(formula = shrimp_survival ~ ammonia_mg_l, data = shrimp_aqua)

Residuals:
      Min        1Q    Median        3Q       Max
-0.075391 -0.017970  0.003363  0.022768  0.064185

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.48982    0.01004  48.798  < 2e-16 ***
ammonia_mg_l -0.14169    0.01990  -7.119 4.83e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02875 on 48 degrees of freedom
Multiple R-squared:  0.5136,    Adjusted R-squared:  0.5035
F-statistic: 50.68 on 1 and 48 DF,  p-value: 4.827e-09

> confint(shrimp_aqua_lm)
                 2.5 %     97.5 %
(Intercept)   0.4696378  0.5100018
ammonia_mg_l -0.1817002 -0.1016706
```

Let’s try it

Hypothesis testing

What’s H0? HA?

What’s the p-value?

How do you interpret it?

Confidence intervals

What interval are you 95% confident contains the population’s coefficient for ammonia?

The mean shrimp survival when ammonia levels are 0 mg l^{-1} could fall in what interval?

If you collected a new data point at ammonia = 0, would you expect it to fall inside or outside the previous range?