Multiple regression

Introduction to Quantitative Ecology
Fall 2018
Chris Sutherland
csutherland@umass.edu

i-Clicker matching

https://tinyurl.com/ycl83wa5

Regression

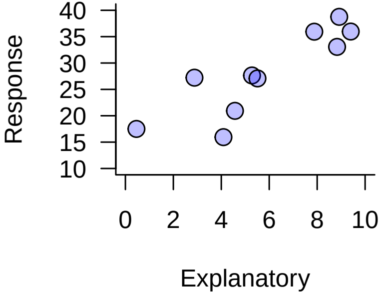
What do you remember?

?

Regression

What we *should* remember about regression:

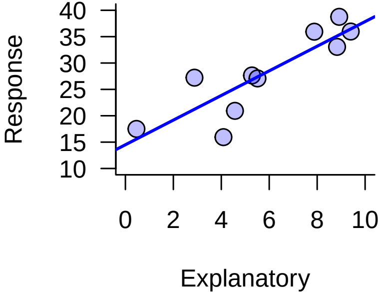
- to find the relationship between two continuous variables



Regression

What we *should* remember about regression:

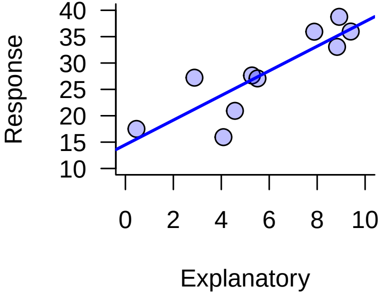
- to find the relationship between two continuous variables
- estimate the correlation coefficient (r)
    - how close are the values to the best fit line



Regression

What we *should* remember about regression:

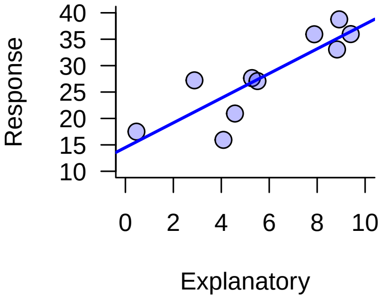
- to find the relationship between two continuous variables
- estimate the correlation coefficient (r)
  - how close are the values to the best fit line
  - 0.86



Regression

What we *should* remember about regression:

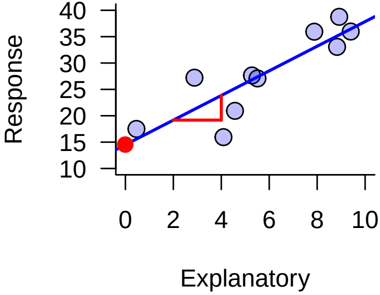
- to find the relationship between two continuous variables
- estimate the correlation coefficient (r)
  - how close are the values to the best fit line
- estimate the parameters of the best fit (straight) line
  - $y = a + bX$



Regression

What we *should* remember about regression:

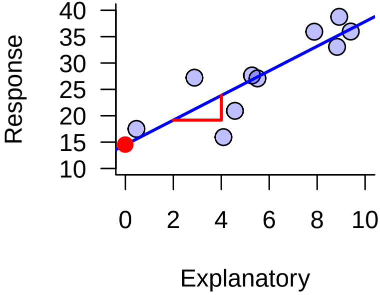
- to find the relationship between two continuous variables
- estimate the correlation coefficient (r)
    - how close are the values to the best fit line
- estimate the parameters of the best fit (straight) line
    - $y = a + bX$
    - y: response variable
    - a: intercept
    - b: slope
    - X: explanatory variable



Regression

What we *should* remember about regression:

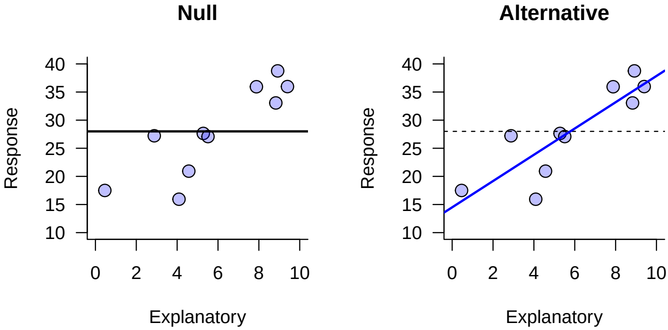
- to find the relationship between two continuous variables
- estimate the correlation coefficient (r)
  - how close are the values to the best fit line
- estimate the parameters of the best fit (straight) line
  - $y = a + bX$
  - y: response variable
  - a: 14.5
  - b: 2.34
  - X: explanatory variable



Regression

Regression analysis provides inference about the slope:

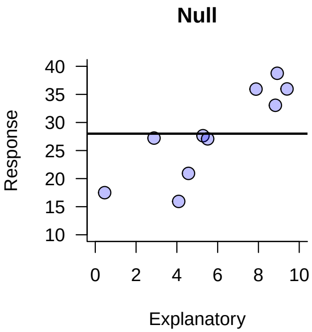
- *null* hypothesis: slope *is not* different from 0
- *alternative* hypothesis: slope *is* different from 0
- how?

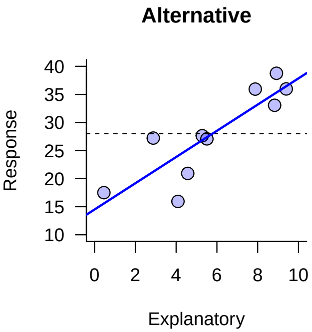


Regression

Regression analysis provides inference about the slope:

- *null* hypothesis: slope *is not* different from 0
- *alternative* hypothesis: slope *is* different from 0
- p -value of the slope





Regression in pRactice

- in algebra
  - $y = a + bX$
- in `R` (a linear model)
  - `lm(y ~ X)`

Regression in pRactice

- in algebra
  - $Response = a + b \times Explanatory$
- in `R` (a linear model)
  - `lm(Response ~ Explanatory, data=df)`

```
df
   Response Explanatory
1     27.22        2.88
2     35.94        7.88
3     15.94        4.09
4     33.06        8.83
5     35.97        9.40
6     17.50        0.46
7     27.64        5.28
8     38.76        8.92
9     27.08        5.51
10    20.93        4.57
```

Regression in pRactice

```r
mod <- lm(Response ~ Explanatory, data = df)
summary(mod)
```

Regression in pRactice

```
mod <- lm(Response ~ Explanatory, data = df)
summary(mod)

Call:
lm(formula = Response ~ Explanatory, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-8.1126 -1.6674  0.2598  2.7585  5.9932

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  14.5011     3.1439   4.612  0.00173 **
Explanatory   2.3353     0.4896   4.770  0.00141 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.325 on 8 degrees of freedom
Multiple R-squared:  0.7399,    Adjusted R-squared:  0.7073
F-statistic: 22.75 on 1 and 8 DF,  p-value: 0.001409
```

Regression in pRactice

```
mod <- lm(Response ~ Explanatory, data = df)
summary(mod)

Call:
lm(formula = Response ~ Explanatory, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-8.1126 -1.6674  0.2598  2.7585  5.9932

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  14.5011     3.1439   4.612  0.00173 **
Explanatory   2.3353     0.4896   4.770  0.00141 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.325 on 8 degrees of freedom
Multiple R-squared:  0.7399,    Adjusted R-squared:  0.7073
F-statistic: 22.75 on 1 and 8 DF,  p-value: 0.001409
```

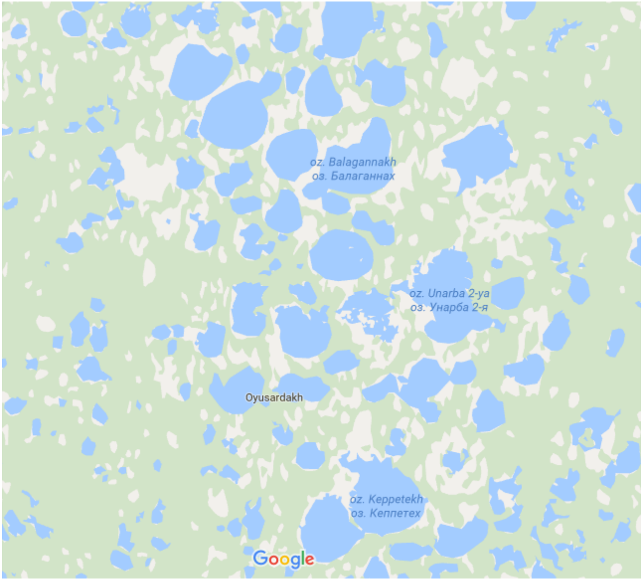
A more complex problem

In typical biological studies:

- rarely collect only a single explanatory variable
- interested in *joint* effects
- interested in *interactive* effects
- we can use *multiple regression*
    - > 1 explanatory variable
    - explanatory variables are continuous

An example - lakes in remote North East Russia

An example - lakes in remote North East Russia



An example - lakes in remote North East Russia

What might influence the number of a certain fish species (abundance) in each of these ponds?

?

An example - lakes in remote North East Russia

What might influence the number of a certain fish species (abundance) in each of these ponds?

- lake size
- pH
- connectivity
- depth
- human activity (e.g., fishing)
- agricultural run-off
- etc…

An example - lakes in remote North East Russia

I am particularly interested in how:

- the size of the lake (`Area`)
- the amount of pesticides in the water (`Pest`)

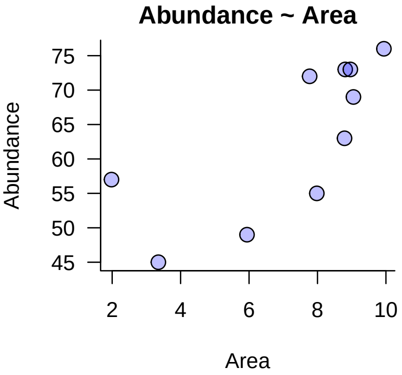
influence the the number of fish counted in a lake (`Abund`)

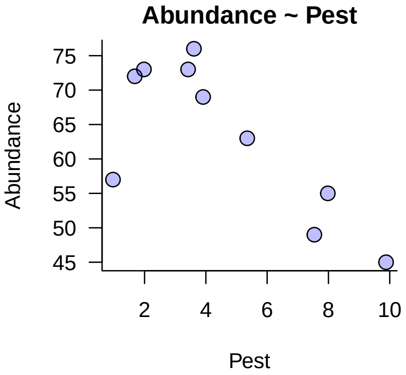
```
fish
   Abund Area Pest
1     55 7.98 7.98
2     49 5.94 7.54
3     69 9.05 3.91
4     73 8.81 3.42
5     76 9.94 3.61
6     73 8.96 1.98
7     63 8.79 5.35
8     57 1.98 0.97
9     45 3.35 9.88
10    72 7.77 1.68
```

An example - lakes in remote North East Russia

To visualize relationships for >1 covariate:

- plot response against each variable independently





Simple regression

Test for an effect of `Area` alone:

```
mod.area <- lm(Abund ~ Area, data = fish)
summary(mod.area)
```

Simple regression

Test for an effect of `Area` alone:

```
mod.area <- lm(Abund ~ Area, data = fish)
summary(mod.area)

Call:
lm(formula = Abund ~ Area, data = fish)

Residuals:
    Min      1Q  Median      3Q     Max
-10.441  -5.806   2.364   4.870  10.156

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  40.7074     7.5614   5.384 0.000659 ***
Area          3.0994     0.9841   3.149 0.013610 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.855 on 8 degrees of freedom
Multiple R-squared:  0.5535,    Adjusted R-squared:  0.4977
F-statistic: 9.919 on 1 and 8 DF,  p-value: 0.01361
```

Simple regression

Test for an effect of `Area` alone:

- significant positive effect
  - $\beta_{\text{Area}} = 3.1$
  - $p = 0.0136099$
  - $R^2 = 0.5$

Simple regression

Test for an effect of `Pest` alone:

```r
mod.pest <- lm(Abund ~ Pest, data = fish)
summary(mod.pest)
```

Simple regression

Test for an effect of `Pest` alone:

```
mod.pest <- lm(Abund ~ Pest, data = fish)
summary(mod.pest)

Call:
lm(formula = Abund ~ Pest, data = fish)

Residuals:
    Min      1Q  Median      3Q     Max
-16.410  -2.534   1.468   3.442   9.950

Coefficients:
            Estimate Std. Error t value    Pr(>|t|)
(Intercept)  76.1147     4.7233  16.115 0.000000221 ***
Pest         -2.7881     0.8704  -3.203      0.0126 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.781 on 8 degrees of freedom
Multiple R-squared:  0.5619,    Adjusted R-squared:  0.5071
F-statistic: 10.26 on 1 and 8 DF,  p-value: 0.01255
```

Simple regression

Test for an effect of `Pest` alone:

- significant negative effect
  - $\beta_{\text{Pest}} = \text{-}2.79$
  - $p = 0.0125513$
  - $R^2 = 0.51$

Simple regression

- significant positive effect of `Area`
  - $\beta_{\text{Area}} = 3.1$
  - $p = 0.0136099$
  - $R^2 = 0.5$
- significant negative effect of `Pest`
  - $\beta_{\text{Pest}} =$ -2.79
  - $p = 0.0125513$
  - $R^2 = 0.51$

Multiple regression

Multiple regression extends the simple linear model:

- to find the relationship between 1 continuous response variable and two or more explanatory variables

Multiple regression

Multiple regression extends the simple linear model:

- to find the relationship between 1 continuous response variable and two or more explanatory variables
- estimate the overall correlation coefficient (R^2)
  - how well the model fits the data

Multiple regression

Multiple regression extends the simple linear model:

- to find the relationship between 1 continuous response variable and two or more explanatory variables
- estimate the overall correlation coefficient (R^2)
  - how well the model fits the data
- estimate the parameters of the best fit relationships
  - $y = a + b_1 X_1 + b_2 X_2 + \ldots$

Multiple regression

Multiple regression extends the simple linear model:

- to find the relationship between 1 continuous response variable and two or more explanatory variables
- estimate the overall correlation coefficient (R^2)
  - how well the model fits the data
- estimate the parameters of the best fit (straight) line
  - $y = a + b_1 X_1 + b_2 X_2 + \ldots$
  - y: response variable
  - a: intercept
  - b_1 b_2: slopes (one for each explanatory variable)
  - X's: explanatory variables

Multiple regression

Multiple regression extends the simple linear model:

- to find the relationship between 1 continuous response variable and two or more explanatory variables
- estimate the overall correlation coefficient (R^2)
  - how well the model fits the data
- estimate the parameters of the best fit (straight) line
  - $y = a + b_1 X_1 + b_2 X_2 + \ldots$
  - b_1 b_2: slopes (one for each explanatory variable)
- are the slopes *significant*?
  - Null: no effect of X_1 on y (i.e., $b_1 = 0$)
  - Null: no effect of X_2 on y (i.e., $b_2 = 0$)
  - test using p -values

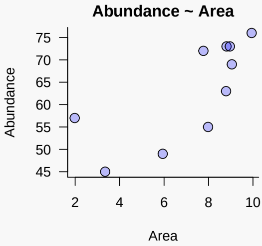
Multiple regression in pRactice

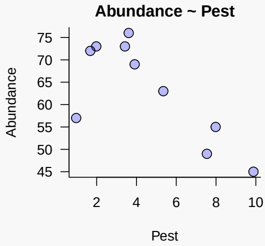
- in algebra
  - $y = a + b_1 X_1 + b_2 X_2$
- in `R`
  - `lm(y ~ X1 + X2)`

Multiple regression in pRactice

- in algebra
  - $Abund = a + b_1 \times Area + b_2 \times Pest$
- in R
  - `lm(Abund ~ Area + Pest, data=fish)`

```
fish
   Abund Area Pest
1     55 7.98 7.98
2     49 5.94 7.54
3     69 9.05 3.91
4     73 8.81 3.42
5     76 9.94 3.61
6     73 8.96 1.98
7     63 8.79 5.35
8     57 1.98 0.97
9     45 3.35 9.88
10    72 7.77 1.68
```





Multiple Regression

Conduct a multiple regression for each:

```
mod.joint <- lm(Abund ~ Area + Pest, data = fish)
summary(mod.joint)
```

Multiple Regression

Conduct a multiple regression for each:

```
mod.joint <- lm(Abund ~ Area + Pest, data = fish)
summary(mod.joint)

Call:
lm(formula = Abund ~ Area + Pest, data = fish)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.9362 -1.9270 -0.7162  2.3124  4.3071 

Coefficients:
            Estimate Std. Error t value   Pr(>|t|)    
(Intercept)  55.1918     3.7078  14.885 0.00000148 ***
Area          2.6037     0.3970   6.558   0.000316 ***
Pest         -2.3503     0.3545  -6.630   0.000296 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.112 on 7 degrees of freedom
Multiple R-squared:  0.9387,    Adjusted R-squared:  0.9212 
F-statistic: 53.57 on 2 and 7 DF,  p-value: 0.00005711
```

Regression results

- significant positive effect of `Area` alone (simple regression)
  - $\beta_{\text{Area}} = 3.1$
  - $p = 0.01361$
  - $R^2 = 0.5$
- significant negative effect of `Pest` alone (simple regression)
  - $\beta_{\text{Pest}} =$ -2.79
  - $p = 0.01255$
  - $R^2 = 0.51$

Regression results

- significant positive effect of `Area` alone (simple regression)
  - $\beta_{\text{Area}} = 3.1$
  - $p = 0.01361$
  - $R^2 = 0.5$
- significant negative effect of `Pest` alone (simple regression)
  - $\beta_{\text{Pest}} = \text{-}2.79$
  - $p = 0.01255$
  - $R^2 = 0.51$
- significant *joint* effects of `Area` & `Pest` (multiple regression)
  - $\beta_{\text{Area}} = 2.6$ ($p = 0.00032$)
  - $\beta_{\text{Pest}} = \text{-}2.35$ ($p = 0.0003$)
  - $R^2 = 0.92$

Group Exercise

Conduct an analysis in `R` to investigate whether the number of voles captured (`Voles`) is influenced by:

- the % vegetation in each habitat patch, `PercVeg`
- the distance to the nearest road , `Dist2Road`
- data: *vole trapping*

1. For each predictor variable (i.e., simple linear regression):
   - Produce a figure to visualize the hypothesis you are testing
   - Fit a linear model
   - How each predictor influences the number of voles captured?
2. Now fit a multiple regression:
   - Produce a 2-panel figure to visualize the hypotheses
   - Fit a 'multiple regression' model
   - How each predictor influences the number of voles captured?
   - Do these results differ to the 'univariate' models
3. (*optional*) Use AIC to pick the 'best' model
   - See page 301 in the book
   - Use the `add1()` function