Welcome to Workshop: Data wrangling & linear models in R

I. Please sign in on the sign in sheet (so I can send you slides & follow up for feedback).

II. If you haven't already, download R and Rstudio, install to your laptop.

III. Download materials you'll need from my website (http://faculty.washington.edu/jhrl/Teaching.html or google Janneke HilleRisLambers at University of Washington – go to Teaching tab, scroll down (zip file under workshop II). Or ask me for a USB stick.

Data wrangling and linear models

I. Goals / Last week

II. Data wrangling

A. Reminder: Projects / scripts

B. ChickenScript_wk2.R (part I); Reading in & examining data, merging and subsetting, defining variables.

C. Nutnet or own data: write a script

III. Linear models

A. Linear models: types & relationships

B. Linear models in R: a quick overview

C. ChickenScript_wk2.R (part II); t-tests, anova, regression / multiple regression, mixed effects models.

IV. Additional Resources

I. Goals / Last week

- What is R; why use R
- Introduction to Rstudio, functions & objects
- Data / project management, coding

These are instructions

Do / look / find this

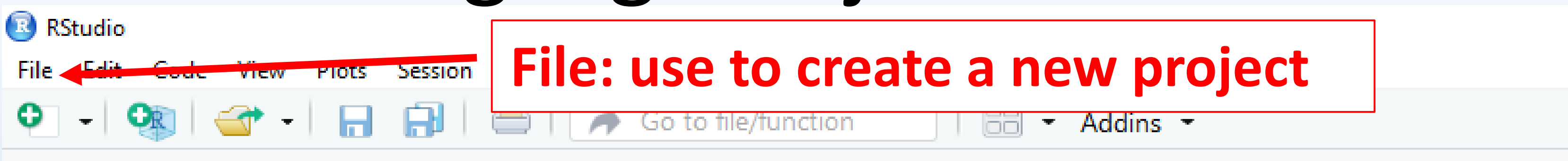
> Type this (but not the >)

This is something useful / important

Instruction:

Open Rstudio

II. Data wrangling: Project for WK 2



Option 1: create a new project

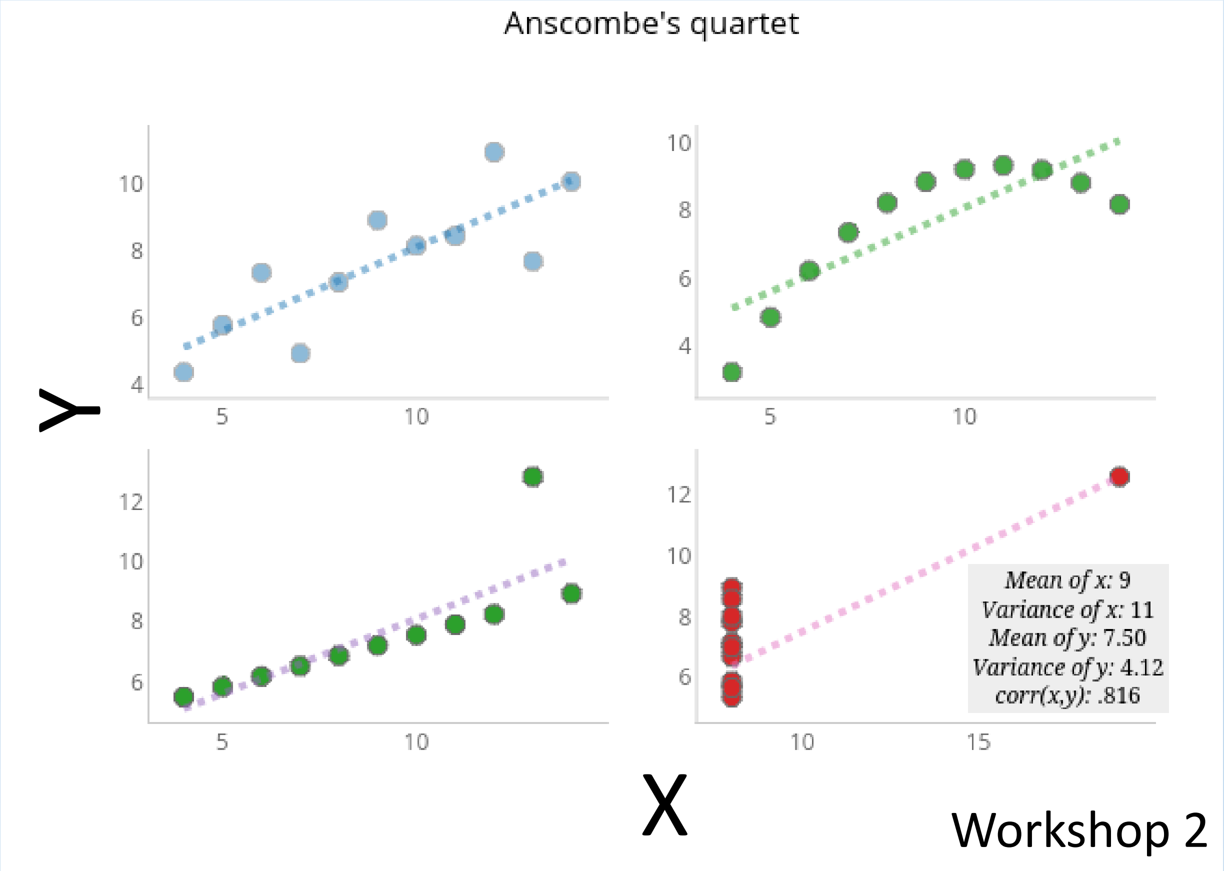
- Go to File / New Project
- Choose New Directory / New Project
- Choose a directory / folder name (e.g. Workshop2) to write in top box
- Choose a location for this directory
- Copy all files for this workshop there

Option 2: use last weeks project

- Browse to directory
- Open your R project (.Rproj file)
- Copy all files for workshop 2 into dir.

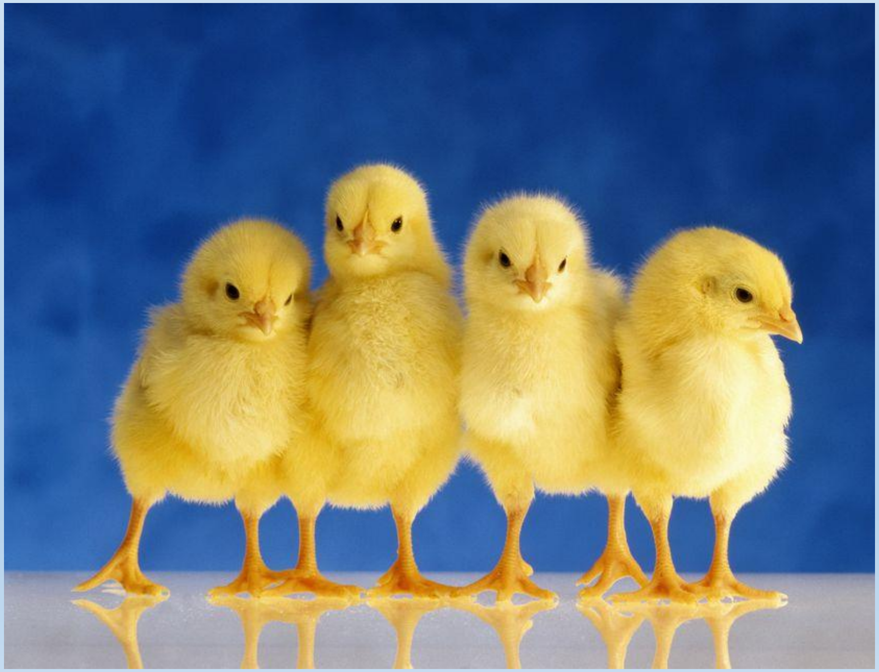
II. Data Wrangling: exploring your data…

- You must explore / examine your data before analyzing it! This includes creating and examining summaries (e.g. mins, means, maxes); logic checks (have I entered all my data?); simple graphs.



II. Data Wrangling: Chickens (intro)

- Data wrangling: read in data, examine data for errors, manipulate data to create summaries, different explanatory variables, exploratory plots.
- You will learn this by running existing code…
- 50 chicks weighed daily for 21 days
- Fed: Soybean, Sunflower, Linseed and Meatmeal



II. Data Wrangling: Chickens (practice)

Instructions

1. Open ChickenScript_wk2.R, and run the code in **Part I** line by line. Try to understand what the code is doing at each step. Note useful functions
2. If you don't have a computer, work with a partner
3. Raise your hand if you have problems.
4. If you finish, try data wrangling for nutnet data (see Nutnet_instrn.pdf)

You can use excel file Rfunctions.xlsx to note useful functions if you'd like

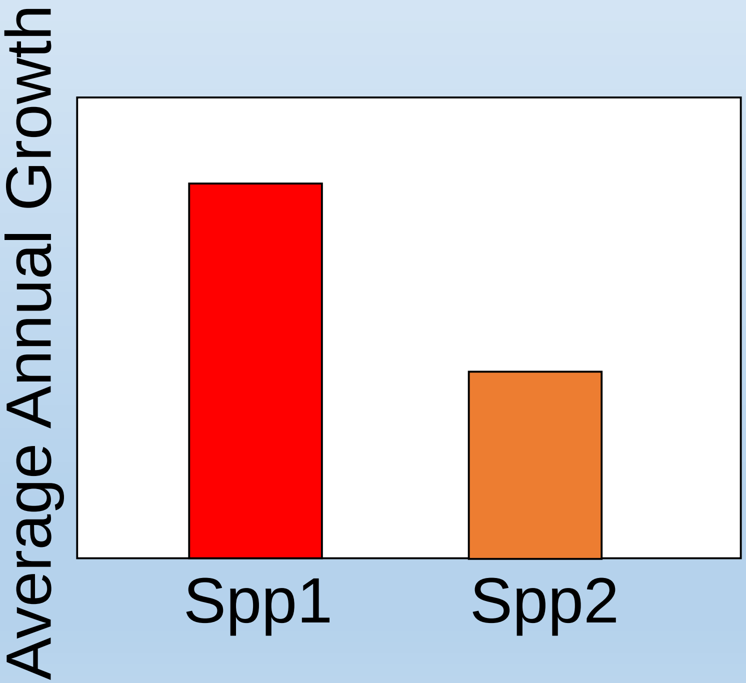
III. Linear models: before you start…

- You should have a (biological) question before you start collecting data, e.g. to test alternative hypotheses. Prediction / parameterization also goals
- Regardless of your goal, you should be able to a) relate the data you collect to your biological question (i.e. hypotheses) OR b) you should be sure that you can estimate / predict with data / stats.
- My suggestion: draw (many) figures representing the patterns you would expect to find in your data under your alternative hypotheses – this can guide you to your statistics.

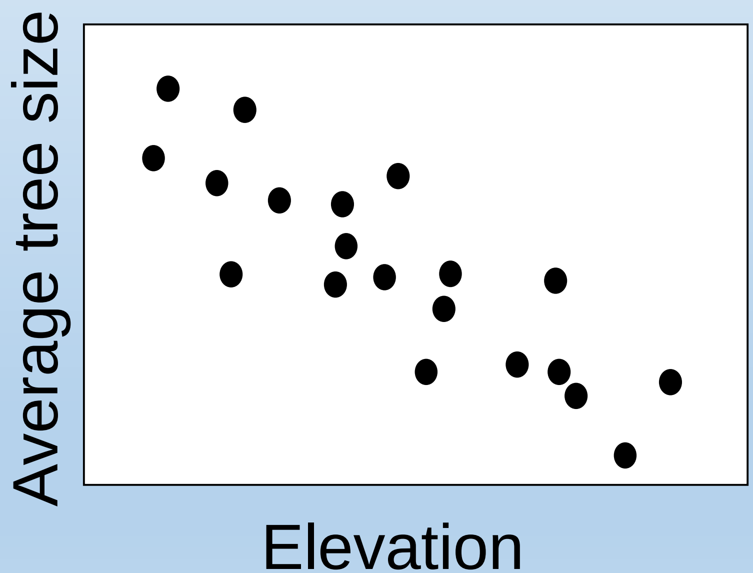
III. Linear models: very brief intro

Many statistical tests boil down to one of two ‘types’ of questions

Is A different from B?



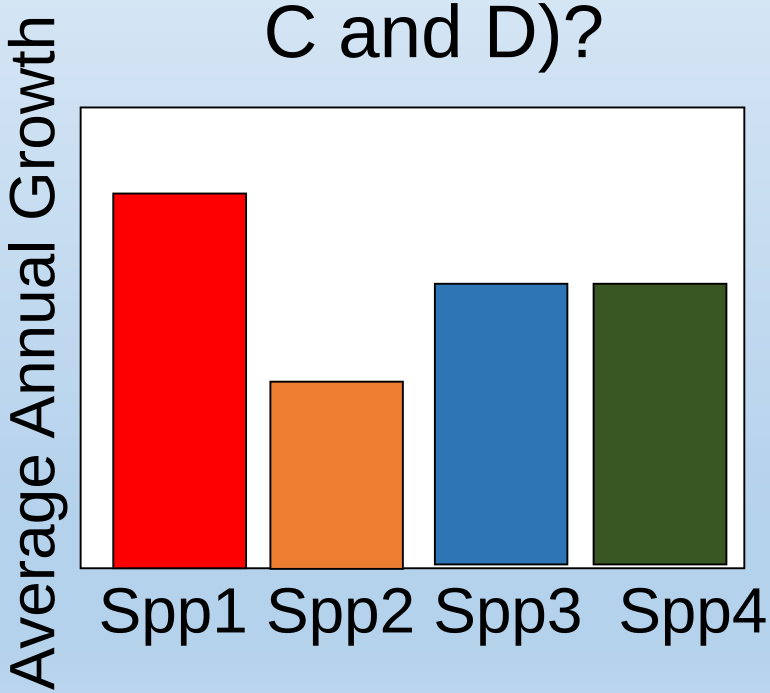
Is A related to B?



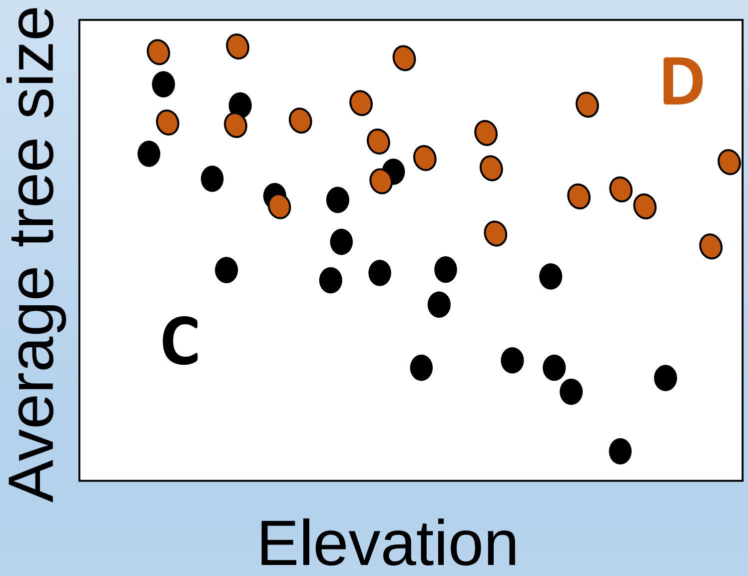
III. Linear models: very brief intro

Many statistical tests boil down to one of two 'types' of questions (or variants thereof)

Is A different from B (and C and D)?



Is A related to B? Does the relationship vary in C vs D?



III. Linear models: very brief intro

What is a linear model? Where predicted values are a linear function of explanatory variables…

Coefficients (e.g. intercept, slope); parameters fitted to data (basis of inference, prediction)

Explanatory variable(s) (you believe influence response variable)

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Response variable (what you are trying to understand)

'errors' (remaining variation in Yi after considering coefficients)

III. Linear models: very brief intro

What is a linear model? Where predicted values are a linear function of explanatory variables…

Your 'ecological model'; a hypothesis for what influences Yi…

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

$$\hat{Y}_i = \boxed{\beta_0 + \beta_1 X_i}$$

Predicted values

$$Y_i - \hat{Y}_i = \varepsilon_i$$

Thus…

III. Linear models: types (relationships)

T-tests, ANOVA's, regressions, generalized linear models, and mixed effects models differ in model fitting / distributions, but are similar in terms of the underlying model…

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

III. Linear models: types (relationships)

T-tests, ANOVA's, regressions, generalized linear models, and mixed effects models differ in model fitting / distributions, but are similar in terms of the underlying model…

THE T-TEST

$$Y_i = \beta_1 + \beta_2 X_i + \varepsilon_i$$

Mean value for category 1, and category 2

Dummy variable (1's and 0's) identifying whether Yi belongs to category 2…

III. Linear models: types (relationships)

T-tests, ANOVA's, regressions, generalized linear models, and mixed effects models differ in model fitting / distributions, but are similar in terms of the underlying model…

THE ANOVA

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} \ldots + \varepsilon_i$$

Mean values for categories 1, 2, 3, etc…

Dummy variables (1's and 0's) identifying whether Yi belongs to categories 2, 3, etc…

Note – 1-way vs 2-way just elaborates on this…

III. Linear models: types (relationships)

T-tests, ANOVA's, regressions, generalized linear models, and mixed effects models differ in model fitting / distributions, but are similar in terms of the underlying model…

THE REGRESSION

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Intercept, slope parameters

Generally, a continuous explanatory variable

III. Linear models: types (relationships)

T-tests, ANOVA's, regressions, generalized linear models, and mixed effects models differ in model fitting / distributions, but are similar in terms of the underlying model…

THE MULTIPLE REGRESSION

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$$

Intercept, multiple slope parameters

Continuous or categorical explanatory variables

Note – ANCOVA is just a multiple regression with 1 continuous variable you don't care about and at least 1 categorical one you do…

III. Linear models: types (relationships)

T-tests, ANOVA's, regressions, generalized linear models, and mixed effects models differ in model fitting / distributions, but are similar in terms of the underlying model…

THE GENERALIZED LINEAR MODEL

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Error distribution can be something different than normal (e.g. Poisson, binomial)

III. Linear models: types (relationships)

T-tests, ANOVA's, regressions, generalized linear models, and mixed effects models differ in model fitting / distributions, but are similar in terms of the underlying model…

Block means

MIXED EFFECTS MODELS

$$Y_{ij} = \beta_0 + \beta_j + \beta_1 X_{ij} + \varepsilon_j + \varepsilon_{ij}$$

Fixed effects: the coefficients & explanatory variables you are interested in

Errors associated with multiple levels – e.g. plot w/in block; and block to block (called random effects)

Note – can get a whole lot more complicated…

III. Linear models: additional jargon

T-tests, ANOVA’s, regressions, generalized linear models, and mixed effects models differ in model fitting / distributions, but are similar in terms of the underlying model…

… But what about t-tests, F-tests, Ordinary least squares, mean squared errors, sums of squares, maximum likelihood?

These are distributions (to test hypotheses), estimation methods, and a few others… Important, but (mostly) don’t affect your ‘biological’ model

III. Linear models: answering questions

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$$

- Estimation: what are the values of B_1 and B_2 (e.g. to plug into simulation models, compare to baseline)?
- Inference: what can be interpreted from coefficients? (e.g. is B_1 positive, implying…?)
- Adequacy: which of multiple models best explain observed data? (e.g. should I include B_1 or B_2 or both)
- How much variation is explained w/ model 1 vs. model 2? (e.g. does climate explain more than competition)
- Prediction: over what range of values can predictions be made for new observations? (e.g. when X1 is __ value, what do I predict Yi is?)

III. Linear models: Chicken Script

Instructions

1. Open ChickenScript_wk2.R, and run the code in **Part II** line by line. Try to understand what the code is doing at each step. Note useful functions.
2. If you don't have a computer, work with a partner
3. Raise your hand if you have problems.
4. If you finish, try statistics with Nutnet data, or Your own? - see Nutnet_instrn.pdf)

Code: t-tests, analysis of variance, linear & multiple regression, and a simple mixed effects model… Other requests? Ask!

III. Your responsibility when writing code…

- You must understand the statistics underlying the code you've written.
- This is true even if (and especially if) you 'pirate' code from someone else / the web (which is perfectly reasonable!).
- You must error proof your code and make sure it is 'reproducible). Run it by someone for review (just like you would a manuscript).
- You must write clear and well commented code that you must be willing to share (increasingly, a requirement for journals).

IV. Further topics: collaborative code

Code sharing / reproducibility: Git and Github

- Git is a free online program that provides version control. Github is the webhosting version of Git.
- Keeps track of all versions of code, associates comments with changes, allows you to resurrect previous versions, and (if coding collaboratively) view changes by coder.
- More and more people are sharing code (in publications) by posting a link to a git repository.

AS A PROJECT DRAGS ON, MY GIT COMMIT MESSAGES GET LESS AND LESS INFORMATIVE.

- Can link to github within Rstudio

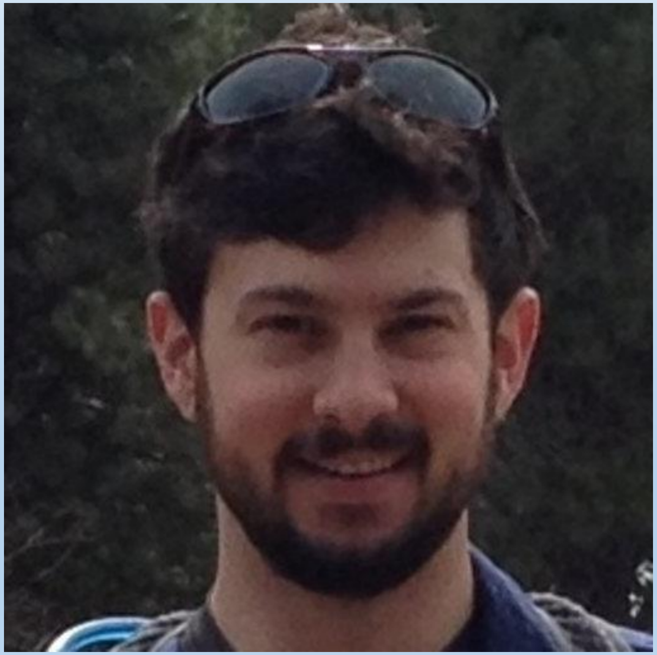
IV. Additional Resources

- Rstudio one page cheatsheets
- Software Carpentry has great workshops (free or virtually free), also online tutorials
- Books: I like Mick Crawley's **The R Book**. Native NZ son Hadley Wickham's book **R for Data Science** is also meant to be good (for data wrangling); he also has a set of packages (check the Tidyverse) that are excellent for munging, merging, data. **Mixed effects models with extensions in Ecology with R** (by Zuur et al) is excellent if you are fitting mixed effects models. **Ben Bolkers' Ecological Models and Data in R** covers linear models, maximum likelihood, and hierarchical Bayesian statistics (a bit)

Any others? Please send them to me!

Acknowledgments

Trevor Branch
UW SAFS – R course (SAFS 552, 553)



Clay Wright
UW Biology, R course (SAFS 552, 553) & Other