C4C: Coding For Conservation

Data and Models

Week 1: June 2022

Goals for this lecture

- To explain what we’re doing here

Goals for this lecture

- To explain what we're doing here
- To define "science"

Goals for this lecture

- To explain what we're doing here
- To define "science"
- To define "data"

Goals for this lecture

- To explain what we're doing here
- To define "science"
- To define "data"
- To define "models"

Goals for this lecture

- To explain what we're doing here
- To define "science"
- To define "data"
- To define "models"
- To introduce many different types of models

Goals for this lecture

- To explain what we're doing here
- To define "science"
- To define "data"
- To define "models"
- To introduce many different types of models
  - Statistical

Goals for this lecture

- To explain what we're doing here
- To define "science"
- To define "data"
- To define "models"
- To introduce many different types of models
  - Statistical
  - Mathematical

All course materials are available at:

coding4conservation.org/syllabus

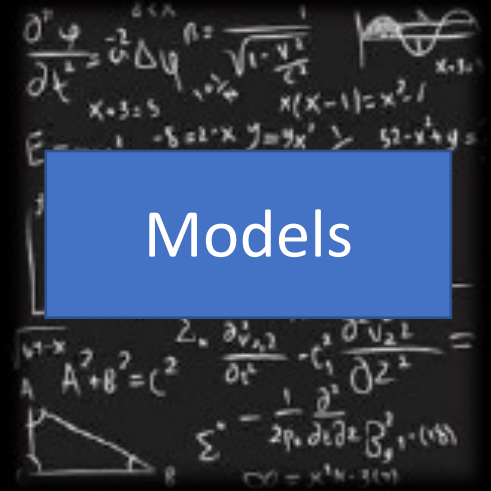
What is science?

the systematic observation of natural events and conditions in order to discover facts about them and to formulate laws and principles based on these facts.

– *Academic Press Dictionary of Science & Technology*

Observations and Laws and Principles

Data and Models



Data and Models

- What is data?

Data and Models

- What is data?
  - Backbone of science

What is science?

the systematic observation of natural events and conditions in order to discover facts about them and to formulate laws and principles based on these facts.

– *Academic Press Dictionary of Science & Technology*

Data vs. Models

- What is data?
  - Backbone of science
  - Evidence to support a claim

Data or not data?

- 19

Data or not data?

- 19
- 19 = total number of fingers and toes

Data or not data?

- 19
- 19 = total number of fingers and toes
- 19 = total number of fingers and toes of Andry Rajoelina

Data or not data?

- 19
- 19 = total number of fingers and toes
- 19 = total number of fingers and toes of Andry Rajoelina
- This is a fact. It becomes data when we use it to support a claim.

There is a negative correlation between the number of years someone has served as president of Madagascar and their total number of fingers and toes.

Data or not data?

- 5, 11, 27

Data or not data?

- 5, 11, 27
- 5, 11, 27 = respective # of children belonging to Mahandry, Ginot, & Tsiry

Data or not data?

- 5, 11, 27
- 5, 11, 27 = respective # of children belonging to Mahandry, Ginot, & Tsiry
- 5, 11, 27 = respective # of children belonging to Mahandry, Ginot, & Tsiry

Mahandry, Ginot, & Tsiry are the names of tenrecs at Duke Lemur Center

Data or not data?

- 5, 11, 27
- 5, 11, 27 = respective # of children belonging to Mahandry, Ginot, & Tsiry.
- 5, 11, 27 = respective # of children belonging to Mahandry, Ginot, & Tsiry.

Mahandry, Ginot, & Tsiry are the names of tenrecs at Duke Lemur Center.

This is a fact.

Data or not data?

- 5, 11, 27
- 5, 11, 27 = respective # of children belonging to Mahandry, Ginot, & Tsiry
- 5, 11, 27 = respective # of children belonging to Mahandry, Ginot, & Tsiry.

Mahandry, Ginot, & Tsiry are the names of tenrecs at Duke Lemur Center.

This is data to support the claim:

Tenrecs have high fecundity rates.

What is data?

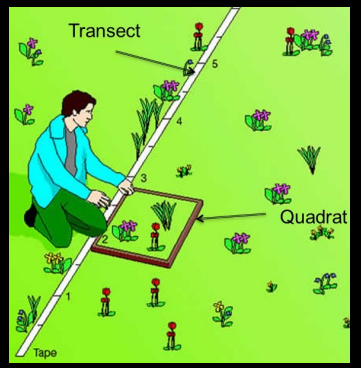
- Backbone of science
- Evidence to support a claim
- A relationship between at least two variables
  - x: explanatory, control, driver, independent variable(s)
  - y: response, dependent variable(s)
- x and y should be clearly defined
  - with respect to the **question!**

Data: Sources of x and y

Data

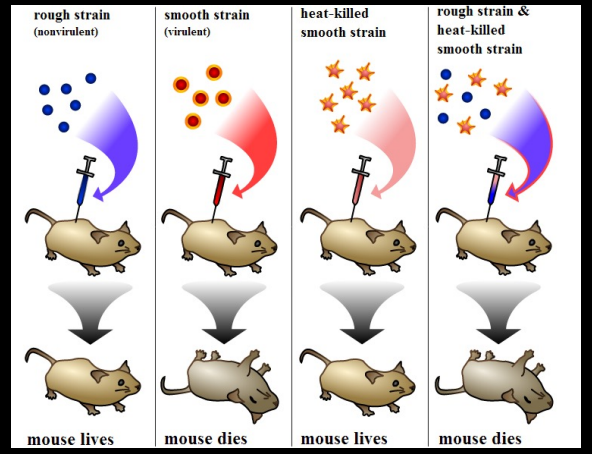
Observational

- Just measure x and y



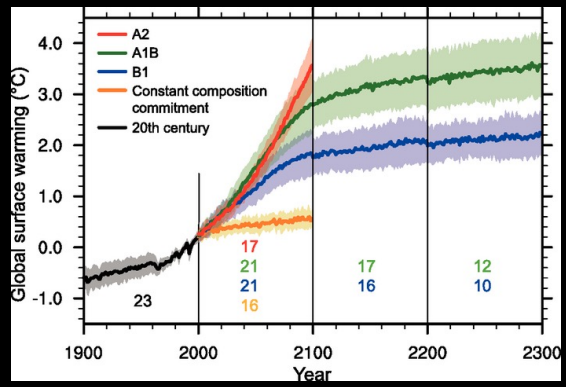
Experimental

- Interfere with x or the relationship between x and y



Simulated

- Create a relationship between x and y



Empirical data

Data: Types

Numerical

Categorical

Data: Types

Numerical

- A variable is numerical when you can transform it with mathematical operation
- Examples?

Categorical

Data: Types

Numerical

- A variable is numerical when you can transform it with mathematical operation
- Examples?
- Integer, real number, multi-dimensional number

Categorical

Data: Types

Numerical

- A variable is numerical when you can transform it with mathematical operation
- Examples?
- Integer, real number, multi-dimensional number

Categorical

- A variable is categorical when it is not numerical but a categorical can be numerical?
- Examples?

Data: Types

Numerical

- A variable is numerical when you can transform it with mathematical operation
- Examples:
- Integer, real number, multi-dimensional number

Categorical

- A variable is categorical when it is not numerical but a categorical can be numerical?
- Examples:
- Colors, (blood) types, species name

Data: Things to consider

- Data acquisition

Data: Things to consider

- Data acquisition
  - Impossible, example?

Data: Things to consider

- Data acquisition
  - Impossible, example?
  - Theoretically possible but practically unfeasible, examples?

Data: Things to consider

- Data acquisition
  - Impossible, example?
  - Theoretically possible but practically unfeasible, examples?
- Data quality and quantity
  - In practice there is always a trade-off

Data: Things to consider

- Data acquisition
  - Impossible, example?
  - Theoretically possible but practically unfeasible, examples?
- Data quality and quantity
  - In practice there is always a trade-off
    - Example: monetary cost, human effort -> power analysis, sampling design etc.

Data: Things to consider

- Data acquisition
  - Impossible, example?
  - Theoretically possible but practically unfeasible, examples?
- Data quality and quantity
  - In practice there is always a trade-off
    - Example: monetary cost, human effort -> power analysis, sampling design etc.
- Reproducibility

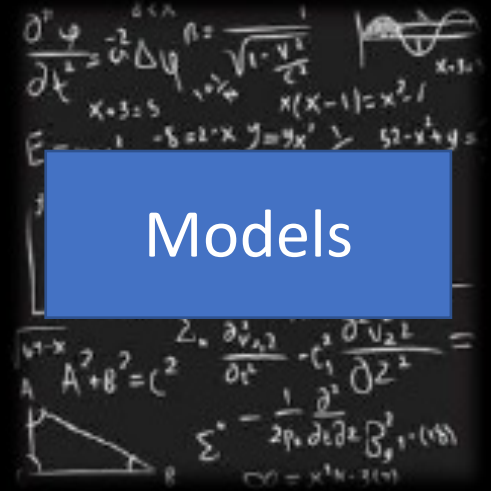
Data: Things to consider

- Data acquisition
  - Impossible, example?
  - Theoretically possible but practically unfeasible, examples?
- Data quality and quantity
  - In practice there is always a trade-off
    - Example: monetary cost, human effort -> power analysis, sampling design etc.
- Reproducibility
- Measurement errors

Data: Things to consider

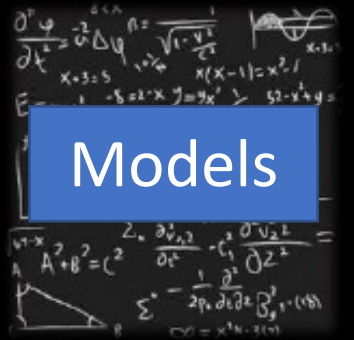
- Data acquisition
  - Impossible, example?
  - Theoretically possible but practically unfeasible, examples?
- Data quality and quantity
  - In practice there is always a trade-off
    - Example: monetary cost, human effort -> power analysis, sampling design etc.
- Reproducibility
- Measurement errors
  - Examples?

Data and Models



Data vs. Models

- What is a model?

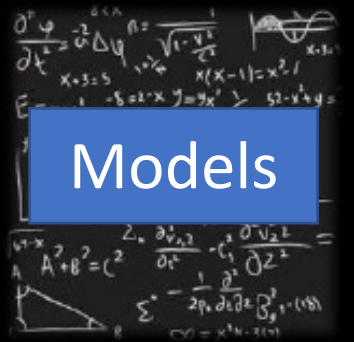


What is science?

the systematic observation of natural events and conditions in order to discover facts about them and to formulate laws and principles based on these facts.

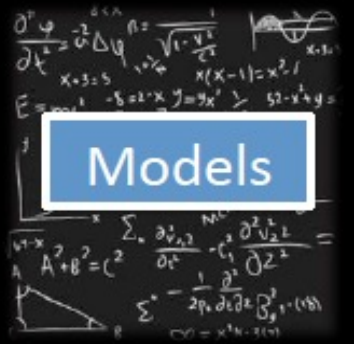
– *Academic Press Dictionary of Science & Technology*

Laws and Principles



- A theory = a declaration to explain a phenomenon
  - Logical and falsifiable
- A model = an abstract representation of a phenomenon
- A hypothesis = a testable declaration that is derived from a theory

Theory, Models, Hypotheses



Theory

Model

Hypothesis

General

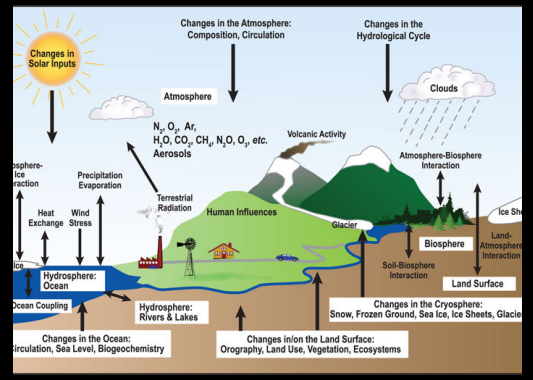
Specific

Models: many types

Human

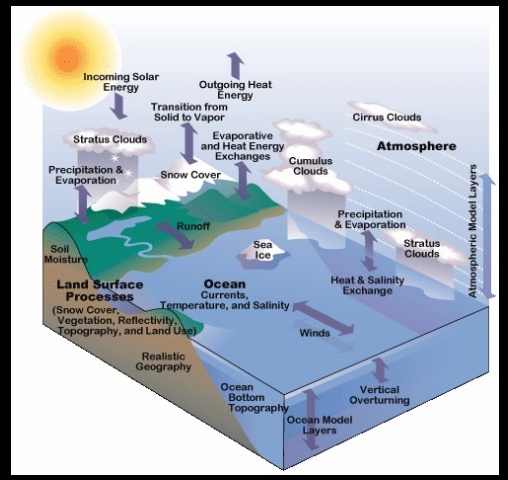
Car

Ecosystem

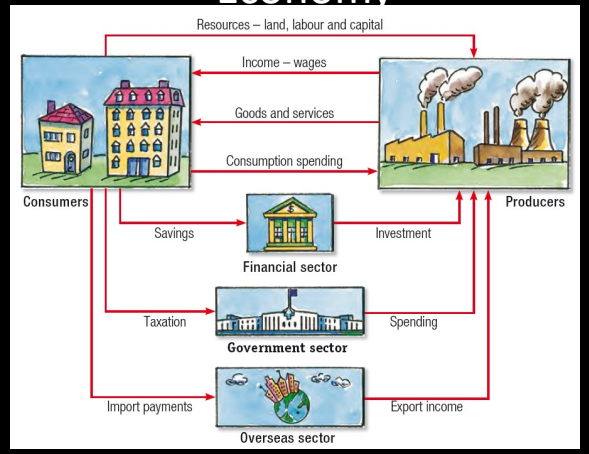


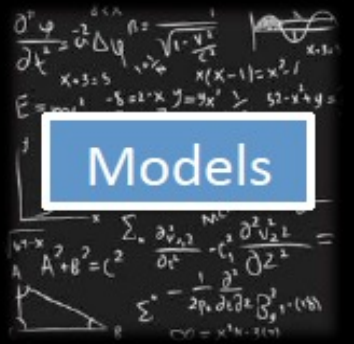
Ecology & Evolution

Climate



Economy

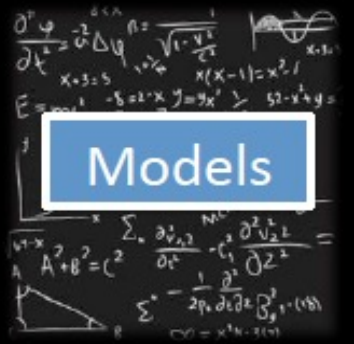




- When you make a model,

  you include the

  elements that you feel are most important
  to explain a phenomenon.



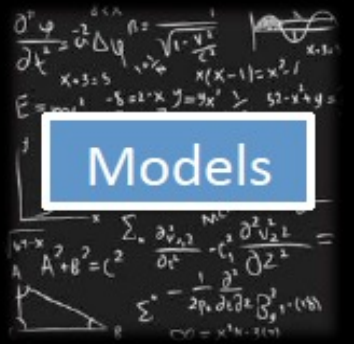
- When you make a model,

  you include the

  elements that you feel are most important
  to explain a phenomenon.

- Generally, we try to make models

  that can reproduce real-world data



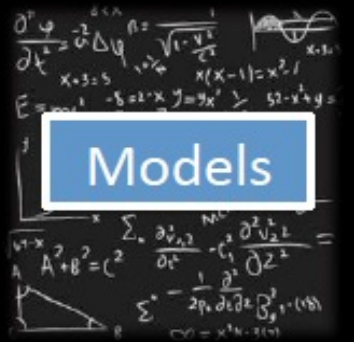
- When you make a model,

  you include the

  elements that you feel are most important
  to explain a phenomenon.
- Generally, we try to make models

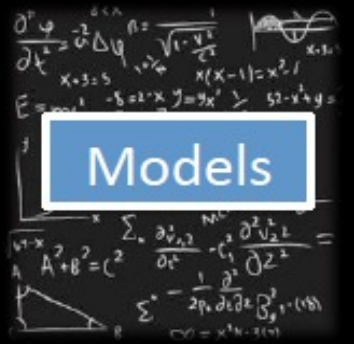
  that can reproduce real-world data
- In C4C, we distinguish between statistical and
  mechanistic models

Statistical vs. Mathematical Model



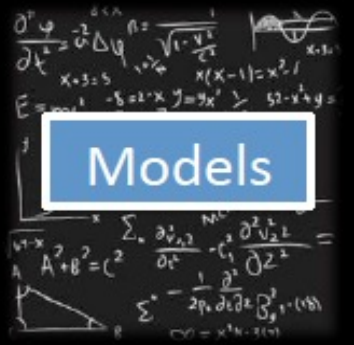
The choice depends on the research question!

Statistical Models



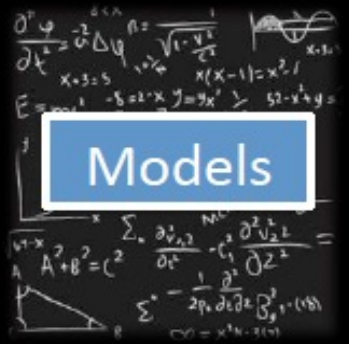
- Goal: To rigorously assess the strength of relationship between x and y
  - Find a significant relationship using a p-value as a measure of relationship strength
  - Statistical models can demonstrate correlations.

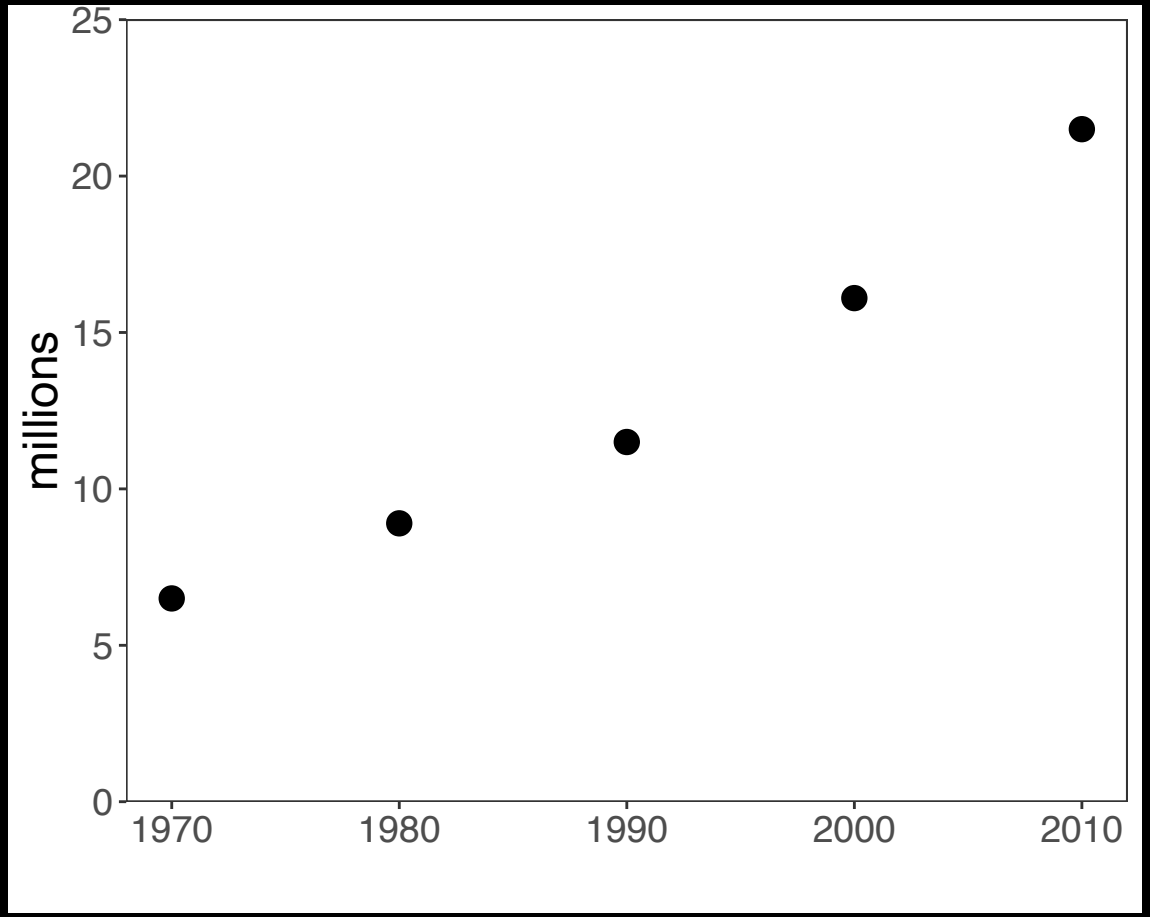
Statistical Models



- Goal: To rigorously assess the strength of relationship between x and y (describe patterns)
  - Find a significant relationship using a p-value as a measure of relationship strength
  - Statistical models can demonstrate correlations.
- Steps:
  1. Formulate a research question
  2. Formulate a hypothesis
  3. Develop a model to demonstrate your hypothesis.
  4. Collect **data (required!!!)**
  5. Evaluate hypothesis with appropriate statistical tools
     - t-test, Chi-square, ANOVA
     - Ordination (PCA)
     - Regression (LM, GLM, GLMM, GAM)

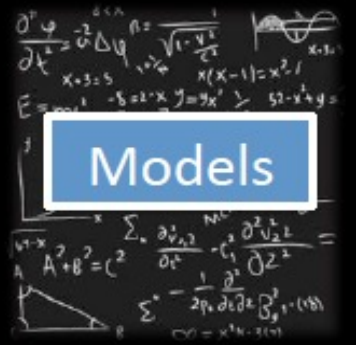
1. Example Question: What is the trajectory Malagasy population size through time?



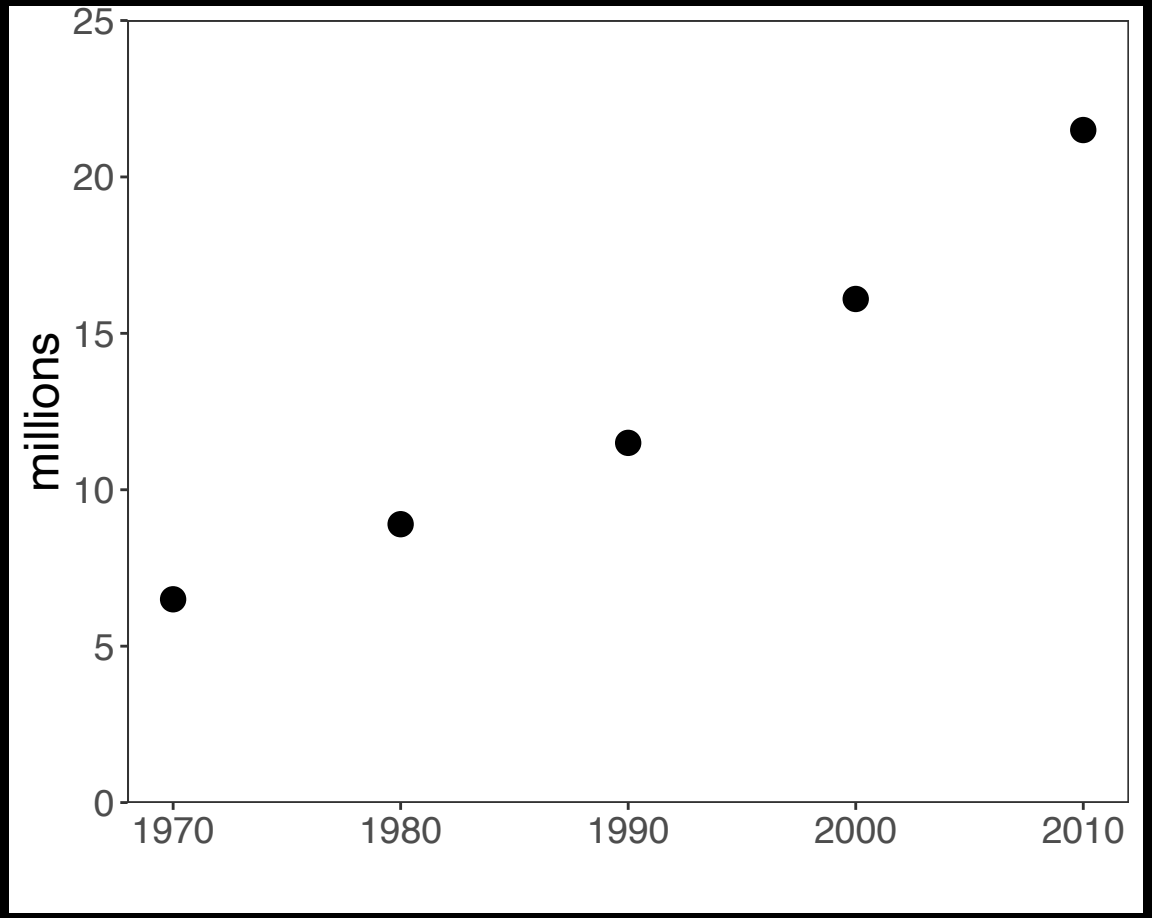


Source: World Bank

1. Example Question: What is the trajectory of Malagasy population size through time?

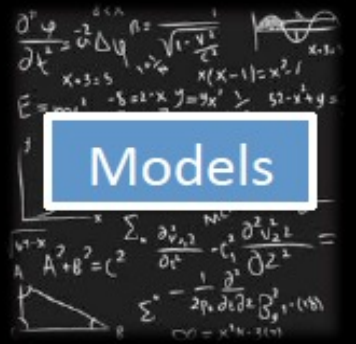


2. Hypothesis: Malagasy population size increases with time



Source: World Bank

1. Example Question: What is the trajectory of Malagasy population size through time?

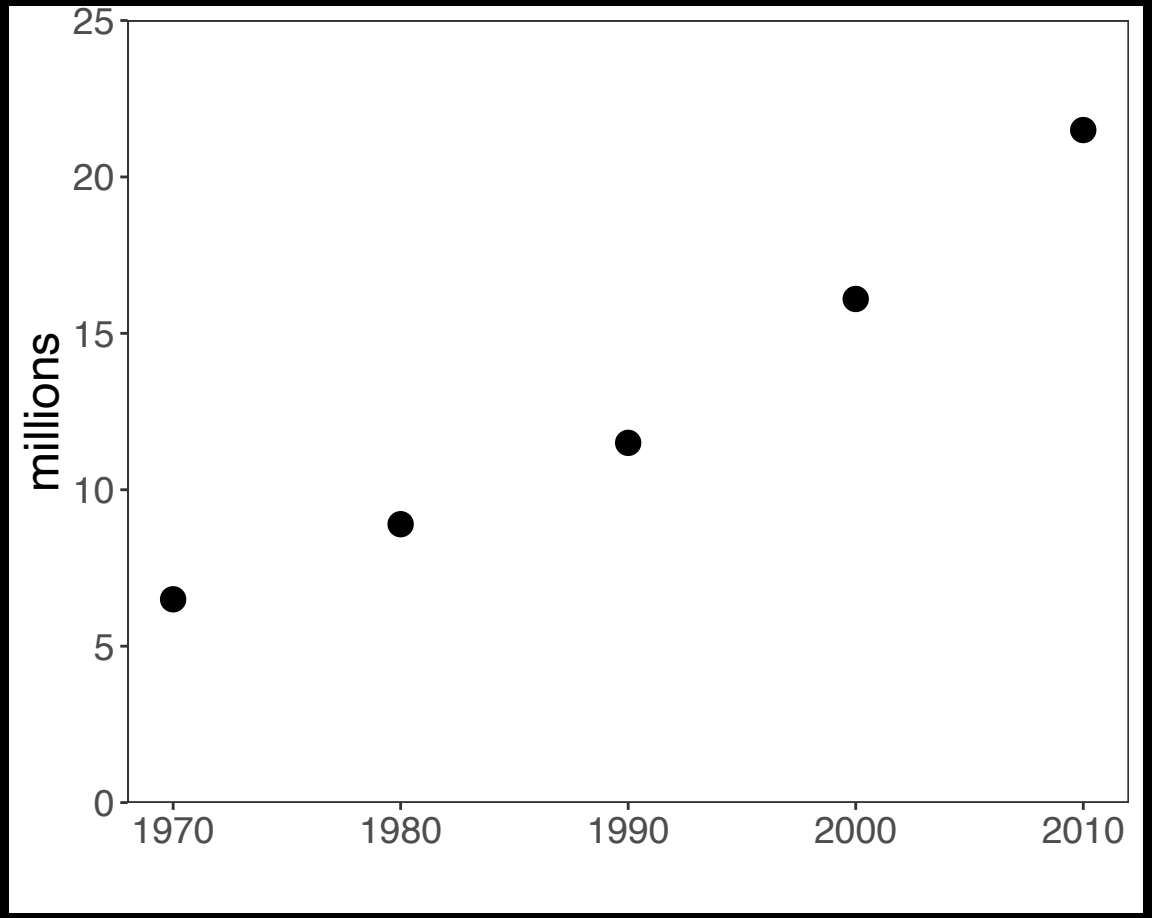


2. Hypothesis: Malagasy population size increases with time

3. Statistical Model:

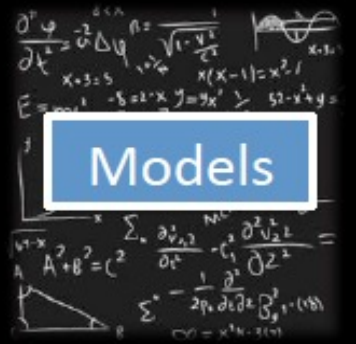
$$y = mx + b$$

Linear Regression



Source: World Bank

1. Example Question: What is the trajectory of Malagasy population size through time?



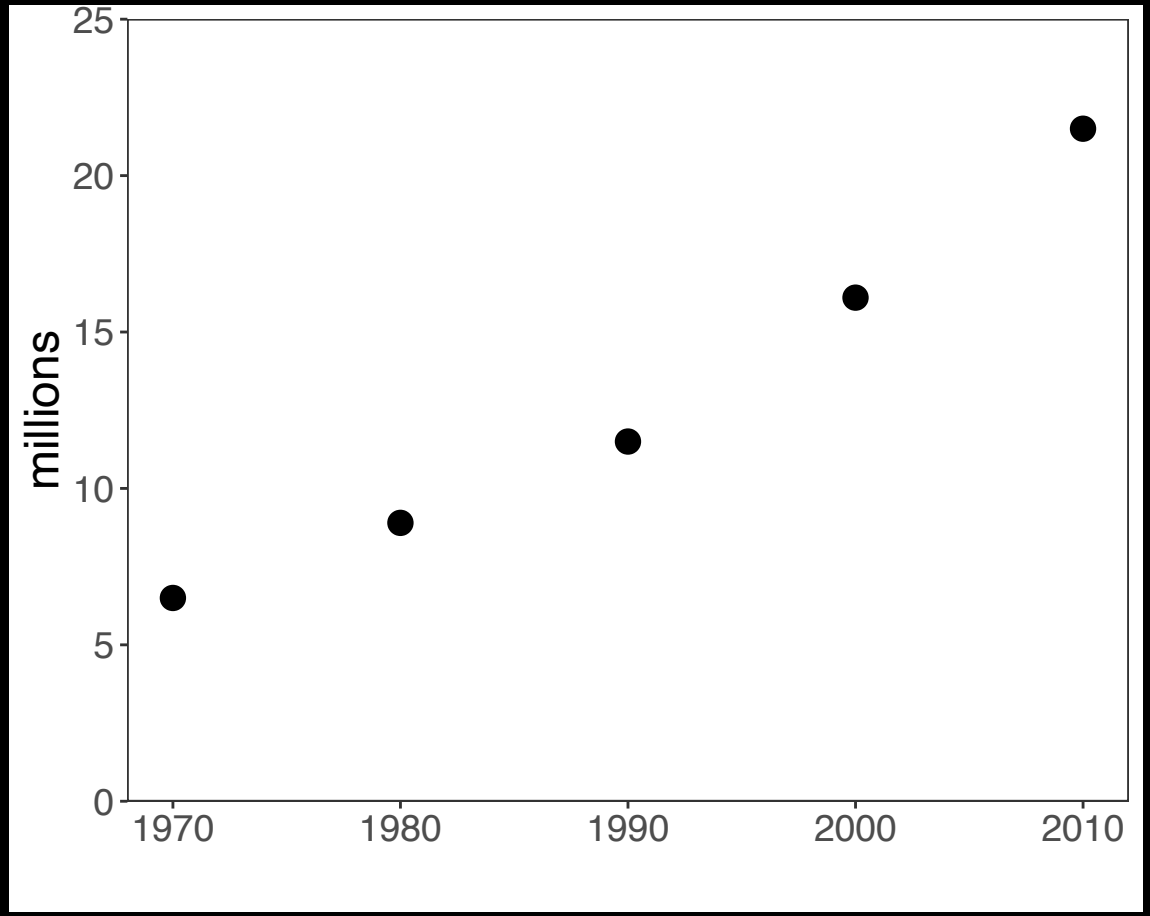
2. Hypothesis: Malagasy population size increases with time

3. Statistical Model:

$$y = mx + b$$

Linear Regression

4. Data:



Source: World Bank

1. Example Question: What is the trajectory of Malagasy population size through time?

2. Hypothesis: Malagasy population size increases with time

3. Statistical Model:

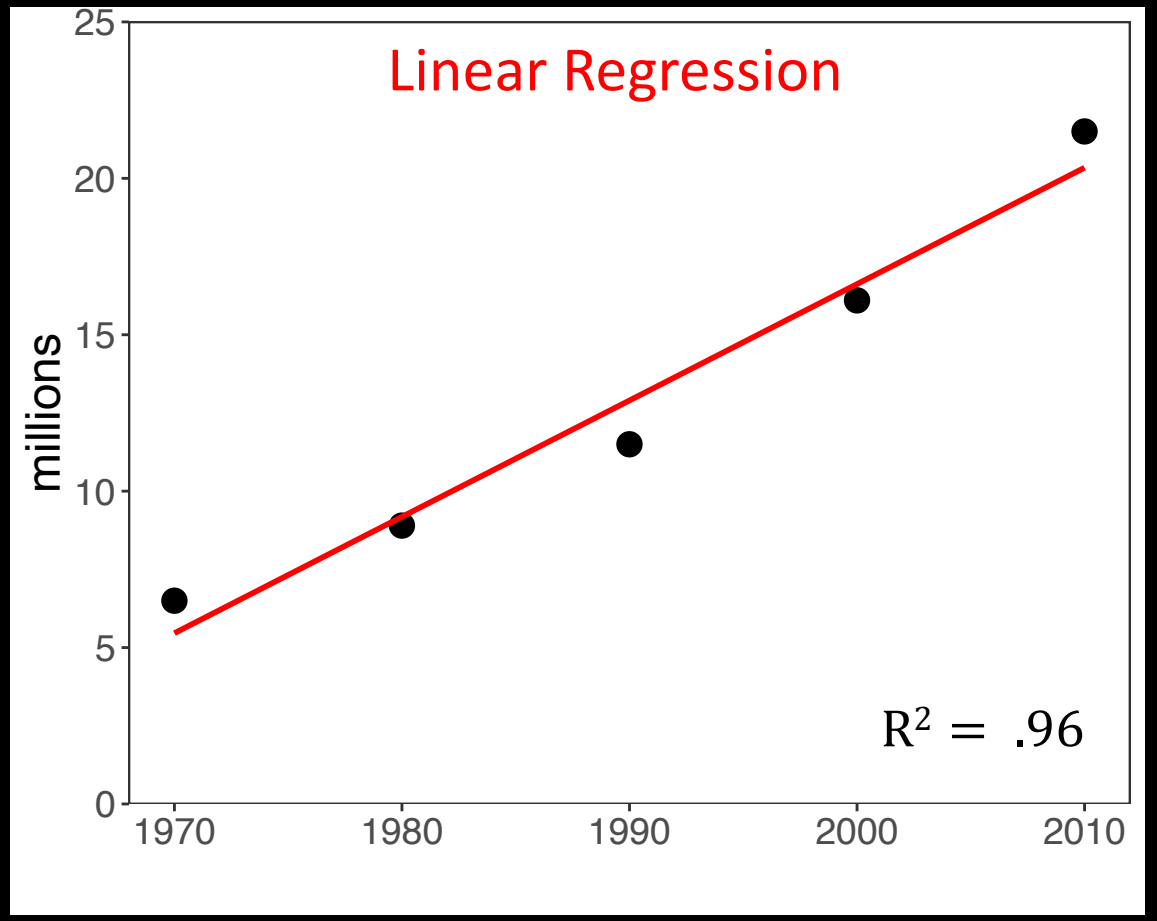
$$y = mx + b$$

5. Evaluation

$$m = .372 \text{ million}$$

$$p = .003$$

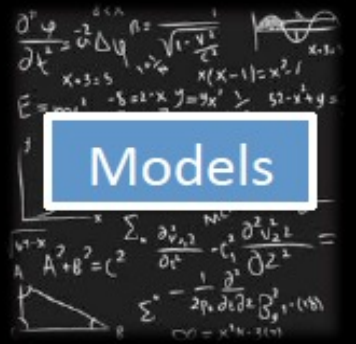
4. Data:



What can we conclude from this fitted model?

Source: World Bank

1. Example Question: What is the trajectory of Malagasy population size through time?



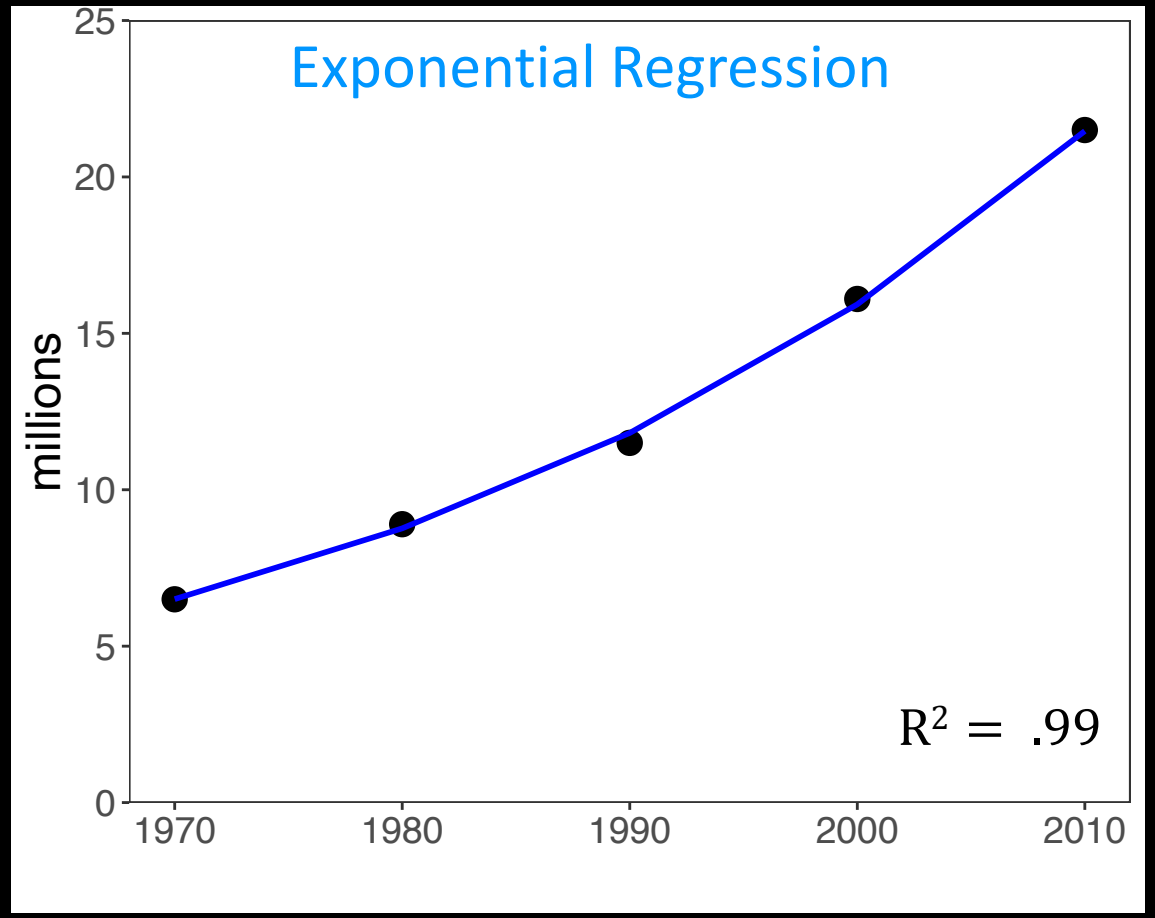
7. Adapt your model and re-evaluate:

$$y = e^{mx+b}$$

Exponential Regression

$m = 0.029$ mil.

$p < .001$

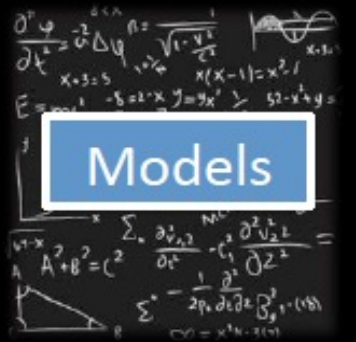


What can we conclude from this fitted model?

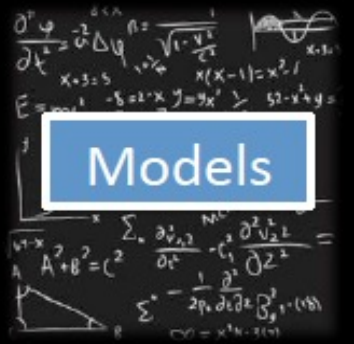
Source: World Bank

Statistical Models: Beware!

- Statistical models and tests are based on specific assumptions
  - data normally distributed
  - y and y independent
  - etc.

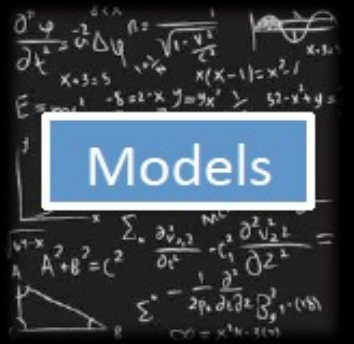


Statistical Models: Beware!



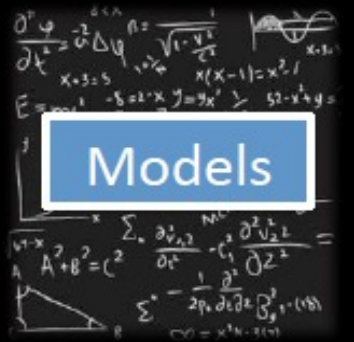
- Statistical models and tests are based on specific assumptions
  - data normally distributed
  - y and y independent
  - etc.
- Assessing a model means you need to make sure the assumptions are not violated.

Statistical Models: Beware!



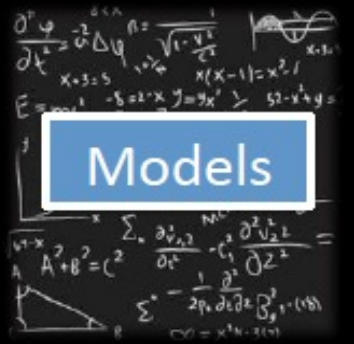
- Statistical models and tests are based on specific assumptions
  - data normally distributed
  - y and y independent
  - etc.
- Assessing a model means you need to make sure the assumptions are not violated.
- There are so many statistical models…

Statistical vs. Mathematical Model



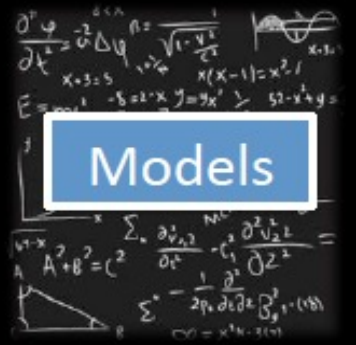
The choice depends on the research question!

Mechanistic Models



- Goal: To demonstrate the processes that underlie a relationship between x and y
  - Find a significant relationship using a p-value as a measure of relationship strength
  - Mechanistic models can demonstrate causation.
- Steps:
  1. Formulate a research question
  2. Formulate a hypothesis
  3. Develop a model to demonstrate your hypothesis.
  4. Collect **data** (for certain questions)
  5. Evaluate the extent to which your model-simulated data matches that from the real world.

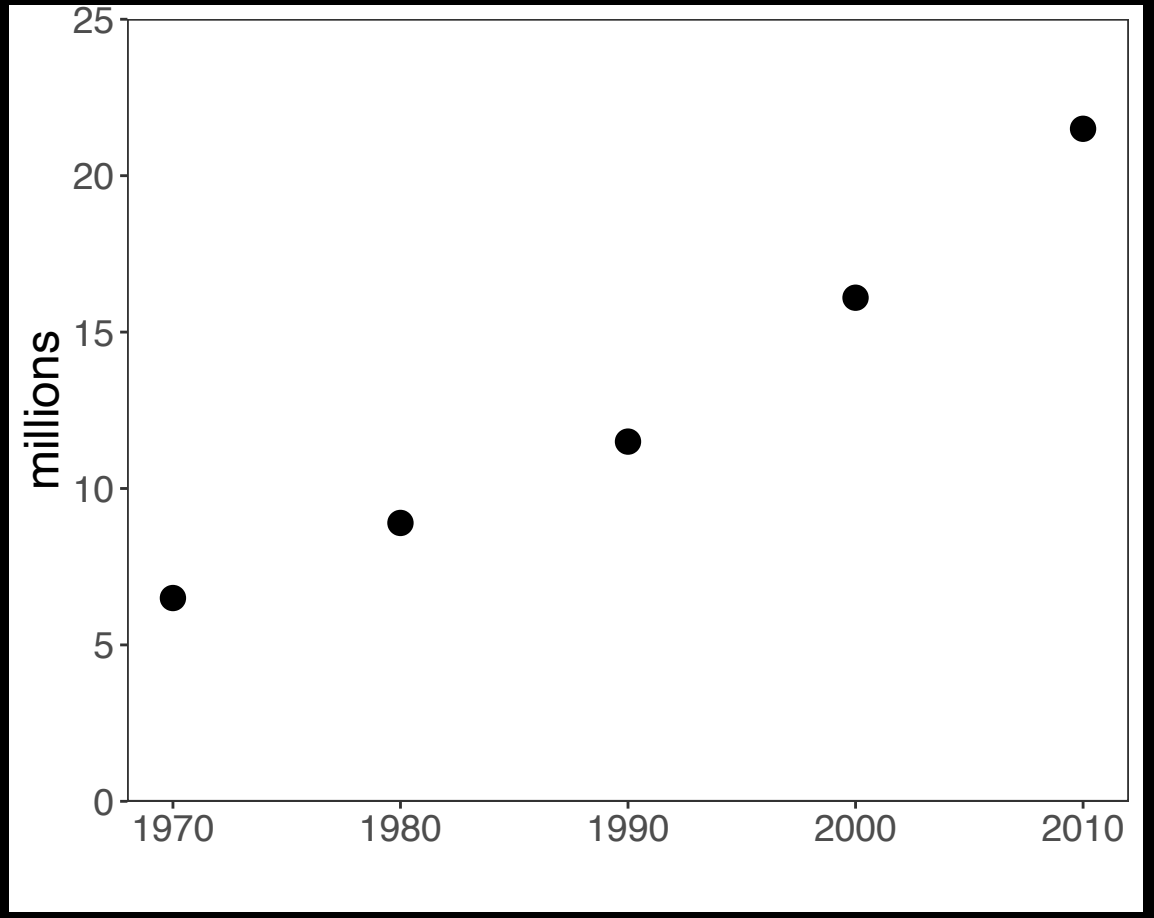
1. Example Question: How does Malagasy population size change with time?



2. Hypothesis: Malagasy population size increases because people are having children.

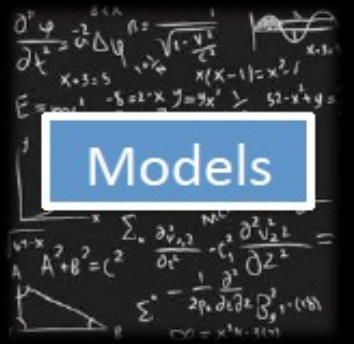
Can you think of an alternative hypothesis?

4. Data:



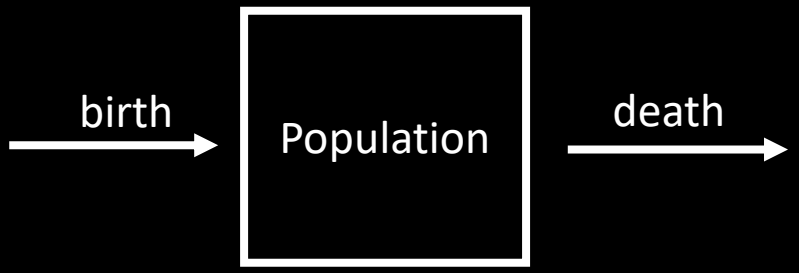
Source: World Bank

1. Example Question: How does Malagasy population size change with time?



2. Hypothesis: Malagasy population size increases because people are having children.

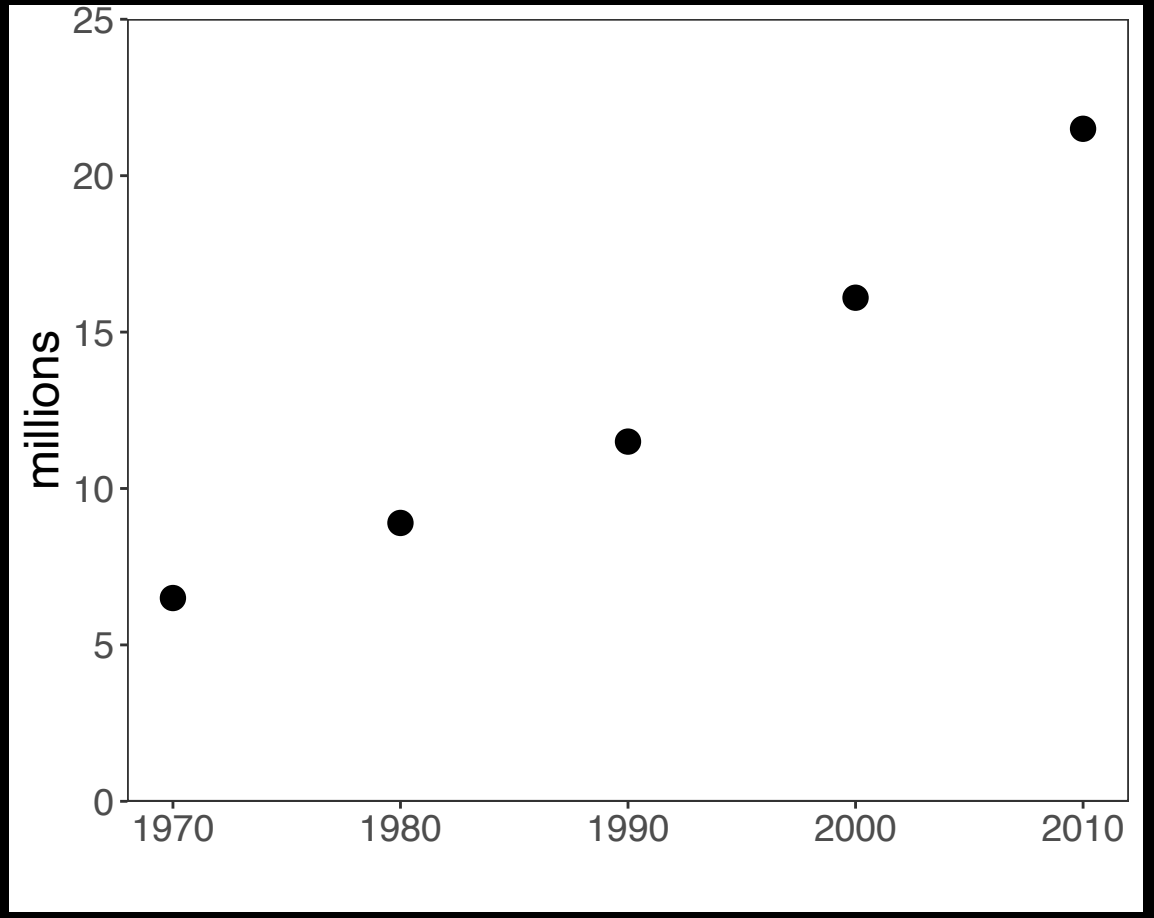
3. Mechanistic Model:



$$P_{t+1} = P_t + b * P_t - d * P_t$$

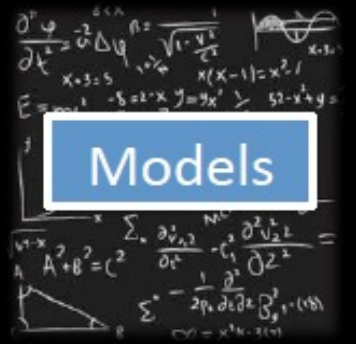
$$P_{t+1} = P_t + r * P_t$$

4. Data:



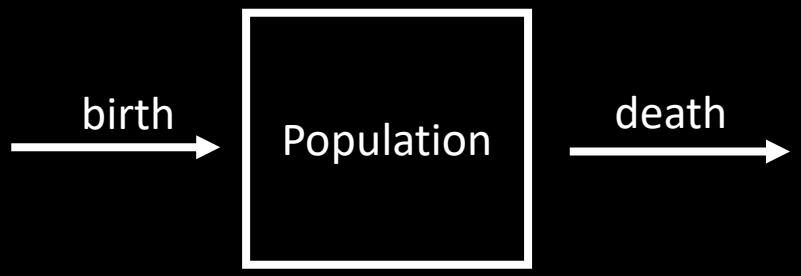
Source: World Bank

1. Example Question: How does Malagasy population size change with time?

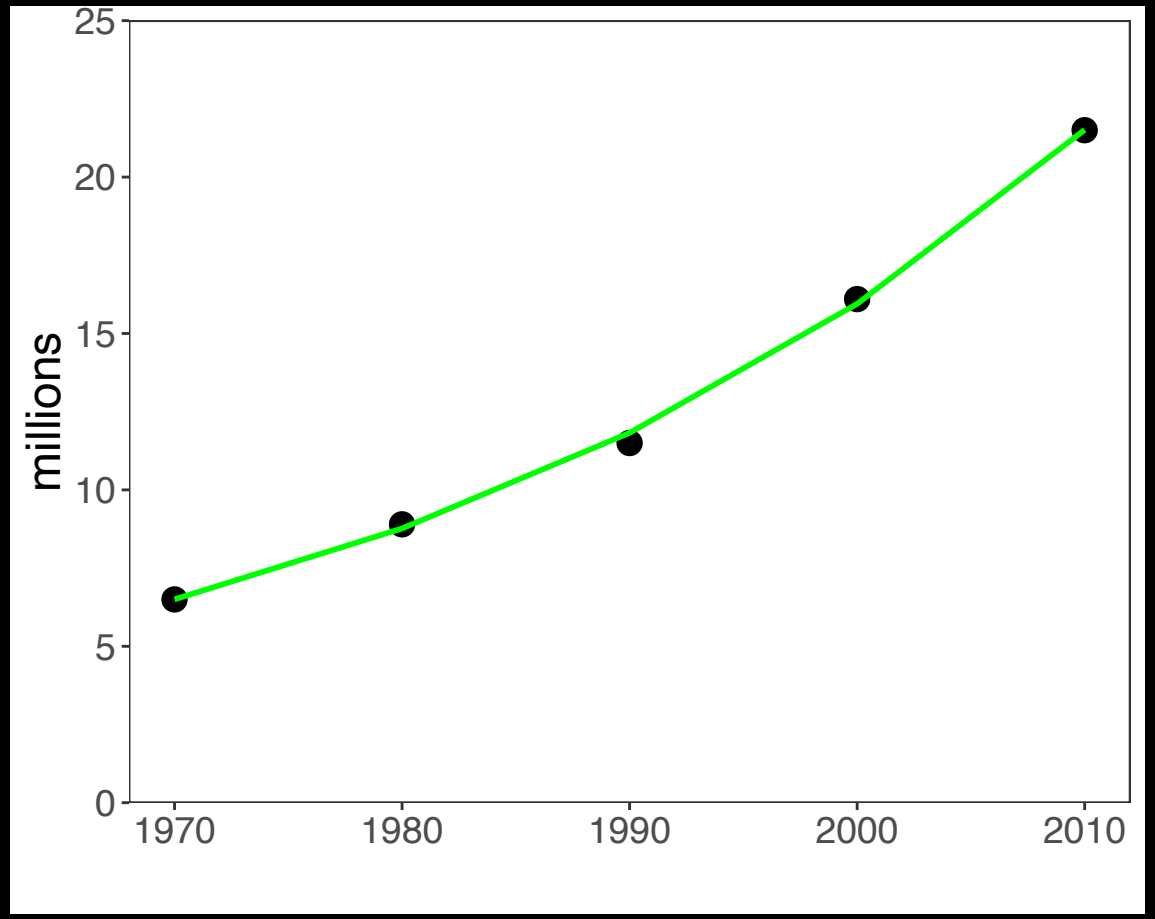


2. Hypothesis: Malagasy population size increases because people are having children.

3. Mechanistic Model:



4. Data:

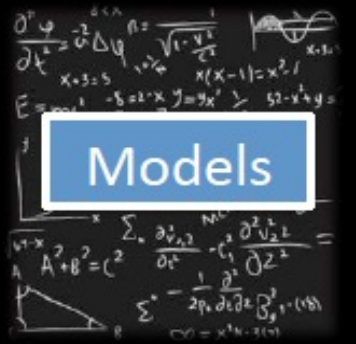


5. Evaluation:

$r = .349$/person/yr

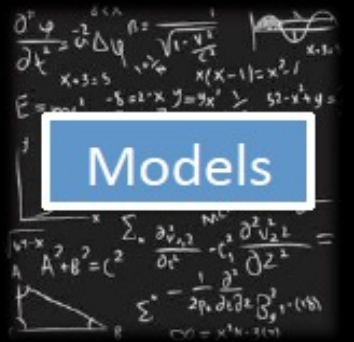
What can we conclude from this fitted model?

Mechanistic Models: Beware!



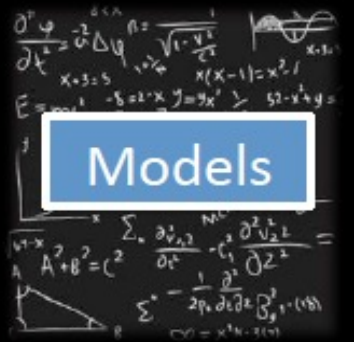
- Parameters used in the mechanistic models sometimes are not measurable!

Mechanistic Models: Beware!



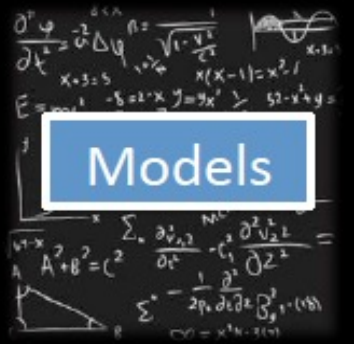
- Parameters used in the mechanistic models sometimes are not measurable!
- Simulations can be computationally intensive

Mechanistic Models: Beware!



- Parameters used in the mechanistic models sometimes are not measurable!
- Simulations can be computationally intensive
- Advances in computational power often inspire development of more complex models which are not necessarily better

Mechanistic Models: Beware!

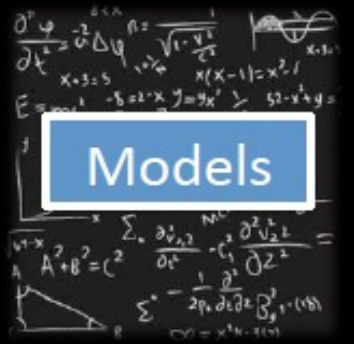


- Parameters used in the mechanistic models sometimes are not measurable!
- Simulations can be computationally intensive
- Advances in computational power often inspire development of more complex models which are not necessarily better

“All models are wrong but some are useful…”

-George Box

Mechanistic Models: Beware!



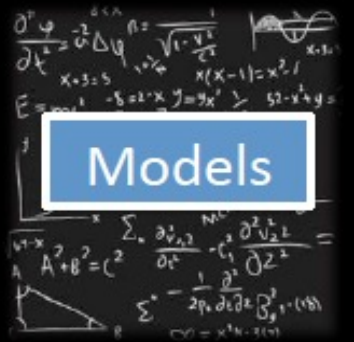
- Parameters used in the mechanistic models sometimes are not measurable!
- Simulations can be computationally intensive
- Advances in computational power often inspire development of more complex models which are not necessarily better

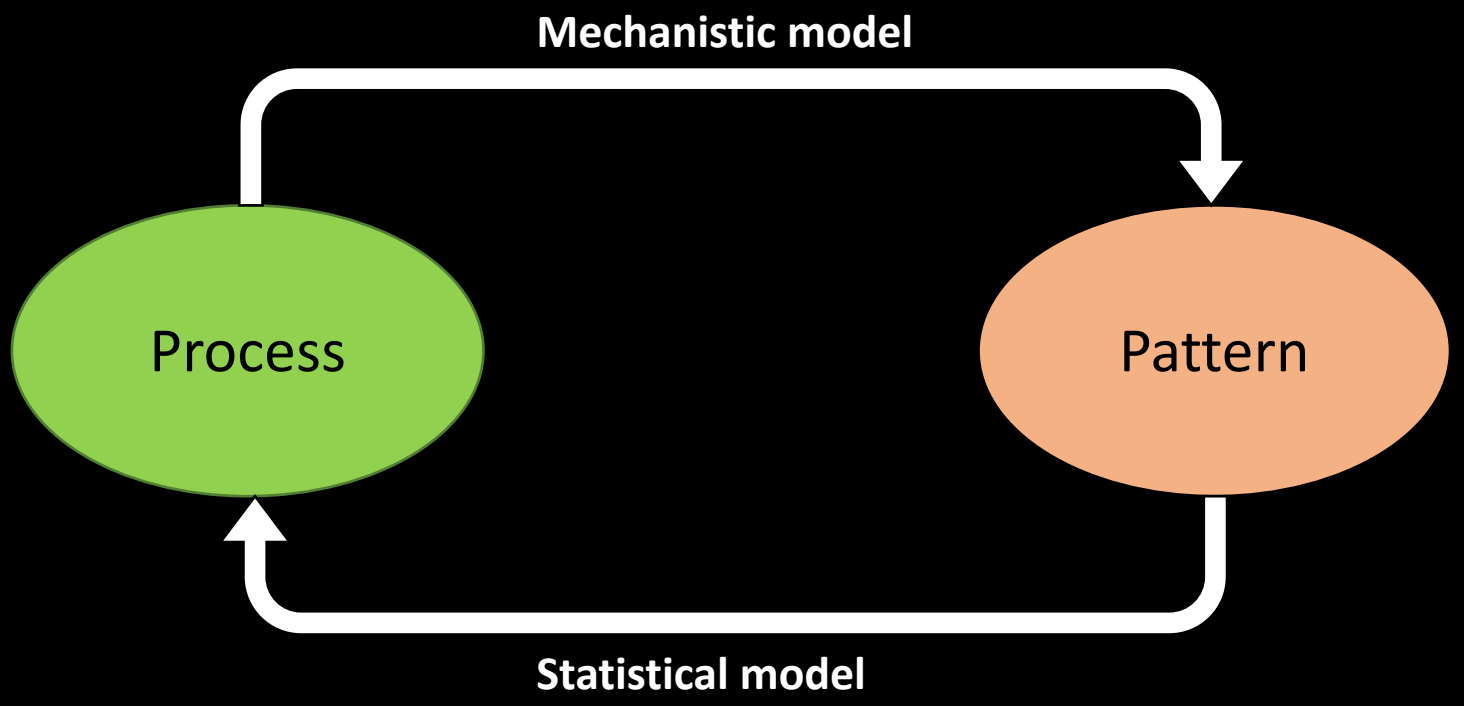
"All models are wrong but some are useful…"

-George Box

We use models to both predict and explain.

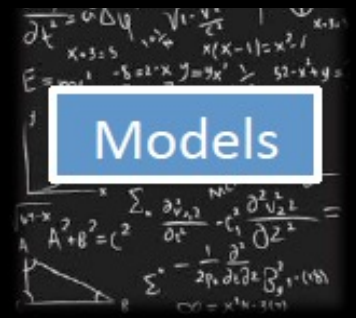
It is ideal when statistical and mechanistic models meet:





A Tool for C4C

- Computer power keeps increasing
- Language/software
  - Fortran, C, C++
  - Julia, Java, Python
  - Matlab, Maple, Mathematica,
  - SAS, SPSS, Stata
- Specific programs
  - Vortex, RAMAS, NetLogo for IBM
  - NicheMapper for physiology, iLand for forest dynamics
  - MaxEnt for species distribution modeling
  - Zonation for reserve selection etc…
- The compromise: R---very powerful for
  - Visualization
  - Data formatting and sorting
  - Statistical analyses
  - Simulation (mechanistic model)



What is R?

- R is a language and environment for statistical computing and graphics. It is used for
    - Data management
    - Statistical analysis
    - Scientific programming and simulation
    - Interfacing with other programs (GIS…)

What is R?

- R is a language and environment for statistical computing and graphics. It is used for
    - Data management
    - Statistical analysis
    - Scientific programming and simulation
    - Interfacing with other programs (GIS…)

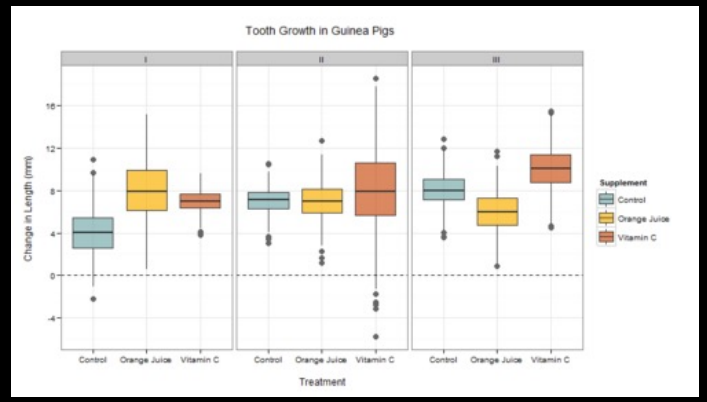
- **Language** because it allows you to communicate flexibly with your computer.

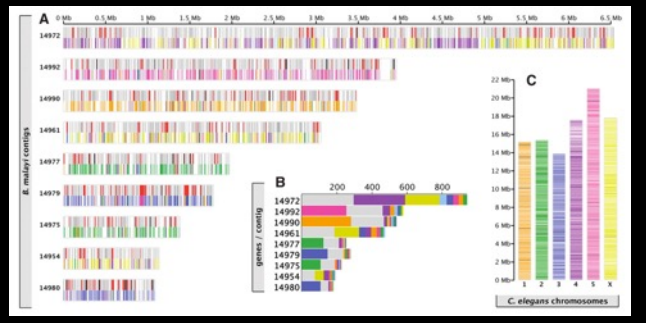
Like any other language:

- Learning R will be easier for some than for others **AND it is okay!!!**
- Learning R takes work and practice

Why use R?

1. **R is free!!!!**
    1. SPSS $99/month
    2. SAS $2,500/year



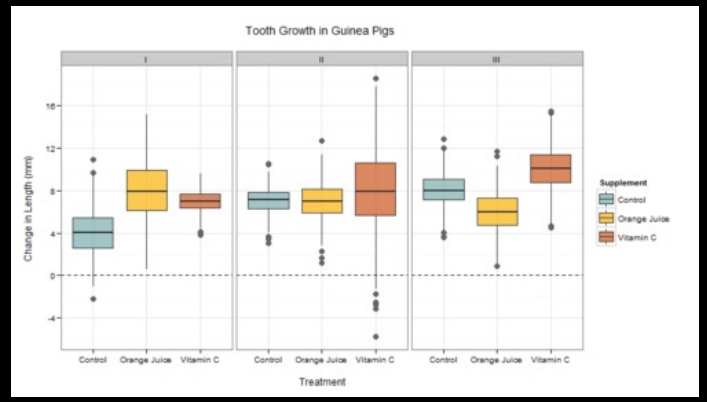


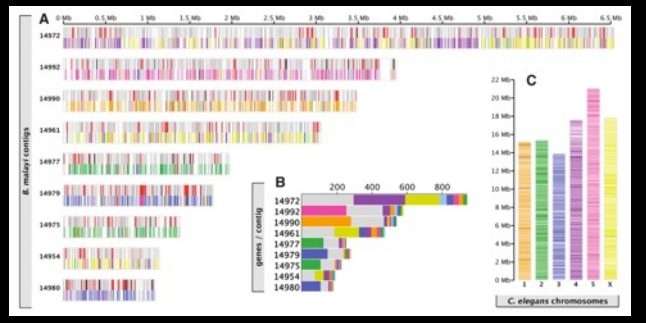
Why use R?

1. **R is free!!!!**
    1. SPSS $99/month
    2. SAS $2,500/year

2. Excellent at making figures

3. Thousands of tools for statistical analysis (packages).

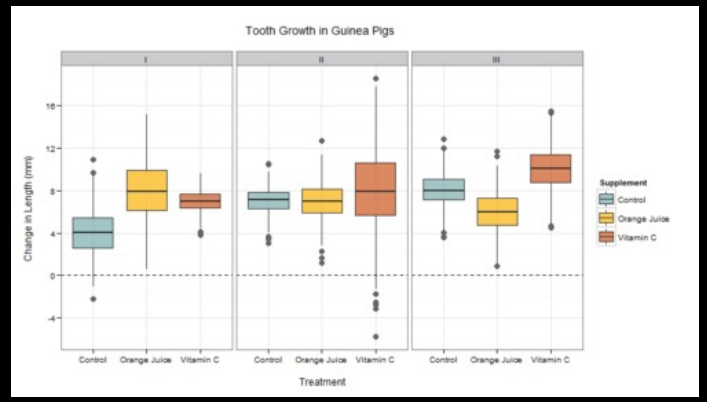




Why use R?

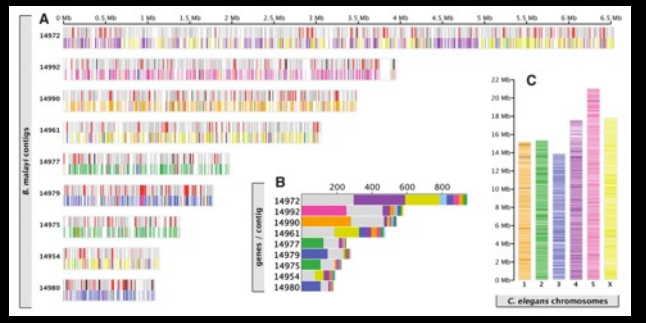
1. **R is free!!!!**
    1. SPSS $99/month
    2. SAS $2,500/year

2. Excellent at making figures



3. Thousands of tools for statistical analysis (packages).

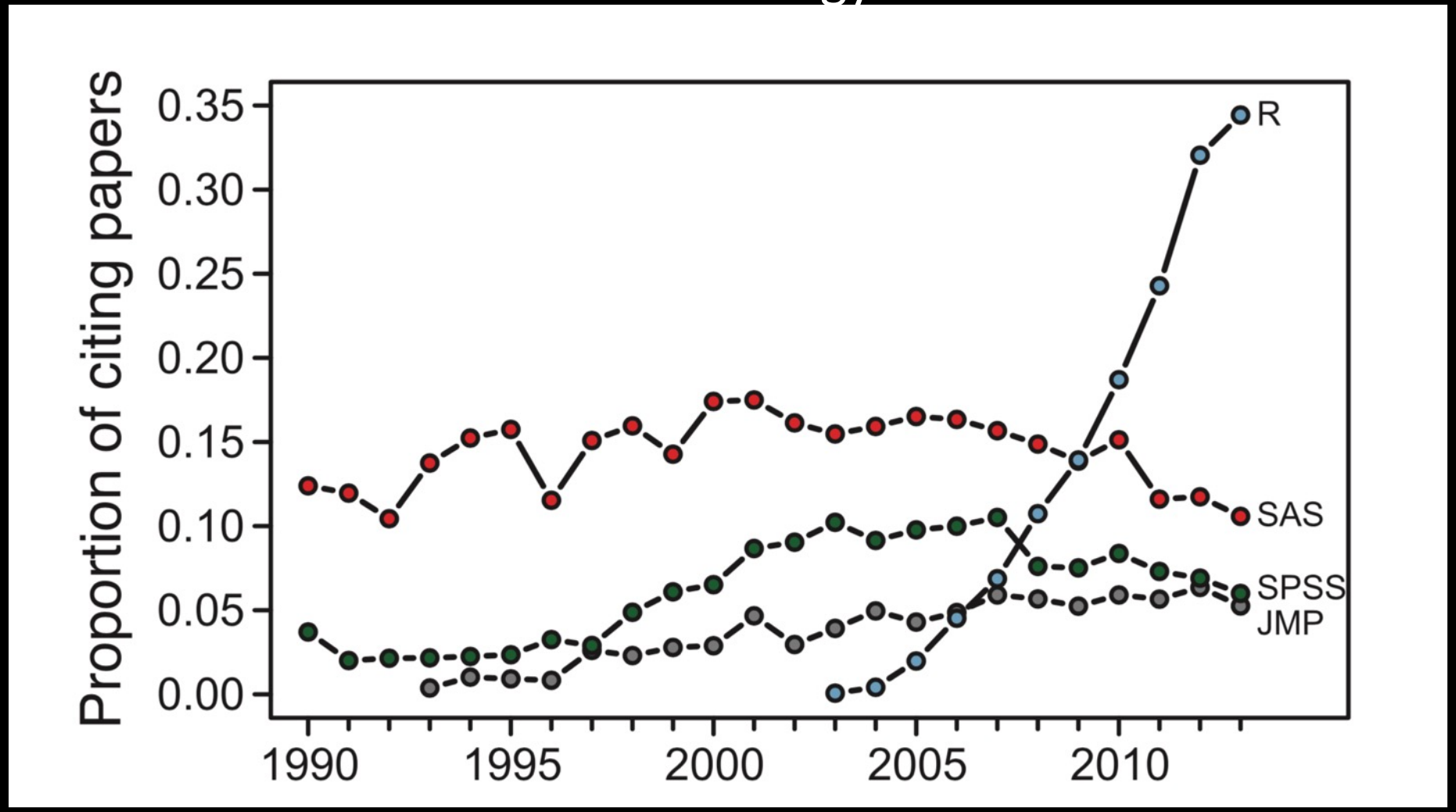
4. Many recently developed tools available immediatly



5. Freedom to develop your own tools

Why use R?

1. Software of reference in ecology



Why use R and how does it work?

The base program is very small (~65 mb)

- Designed to have task-specific packages downloaded and added to it. **There is probably a package that is designed to do the analysis that you want to do**

- A package is a collections of functions, data, and help files generally centered around certain themes of analyses.

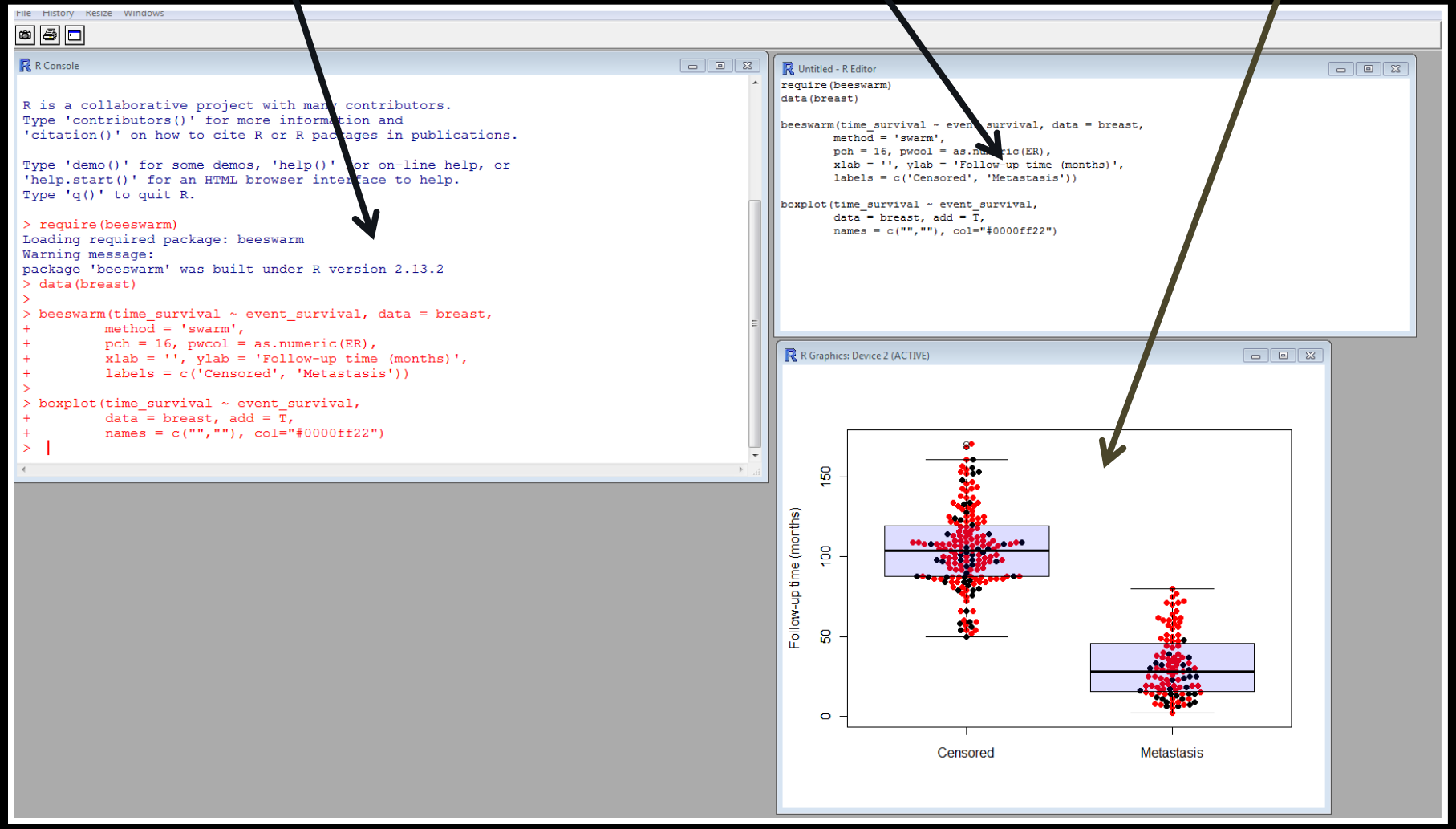
- 10,000+ packages are currently available to download (you will never need most of these)

Your environment in R

Console

Editor

Graphics

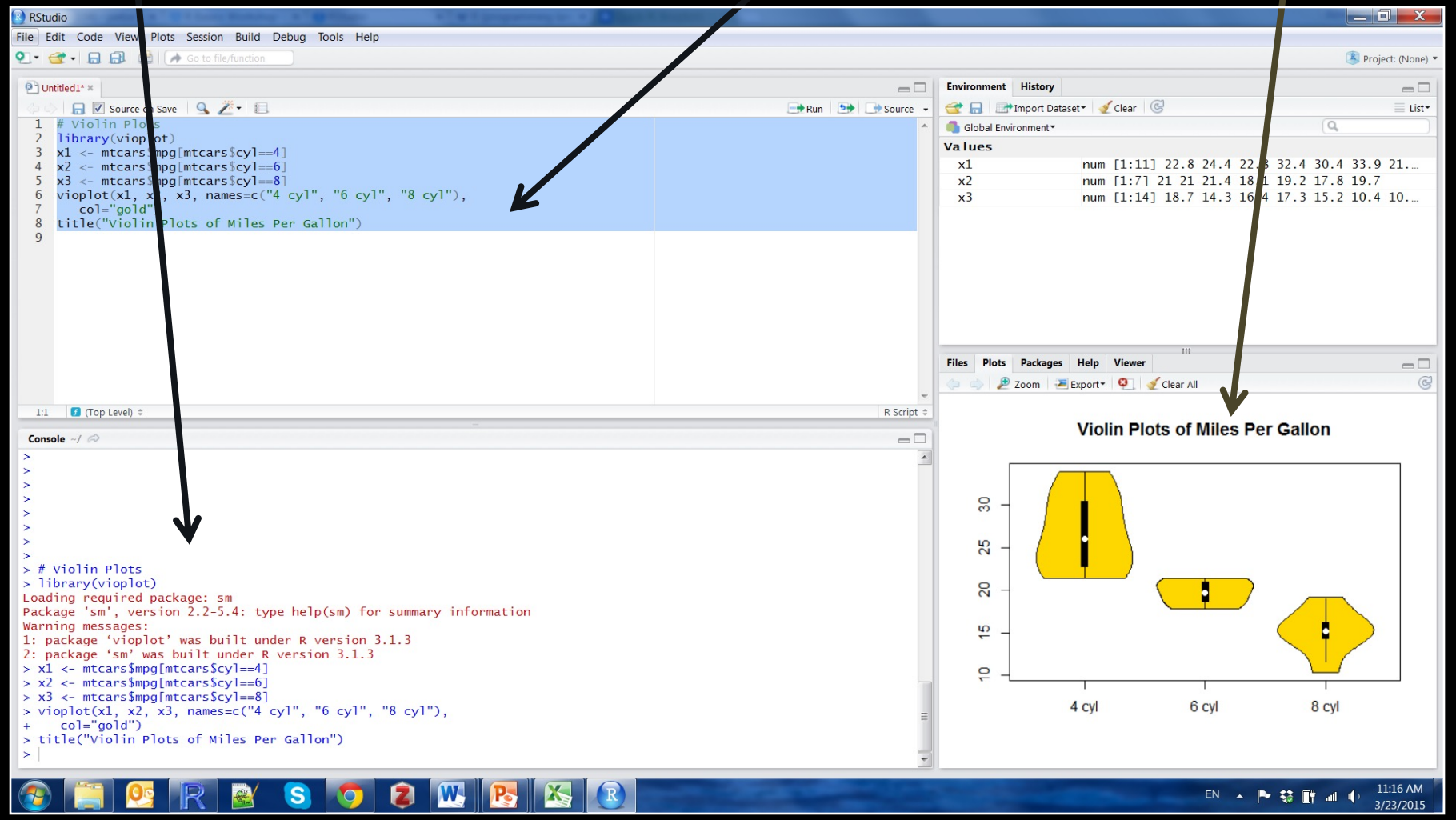


Main windows in R Studio

Console

Editor

Graphics



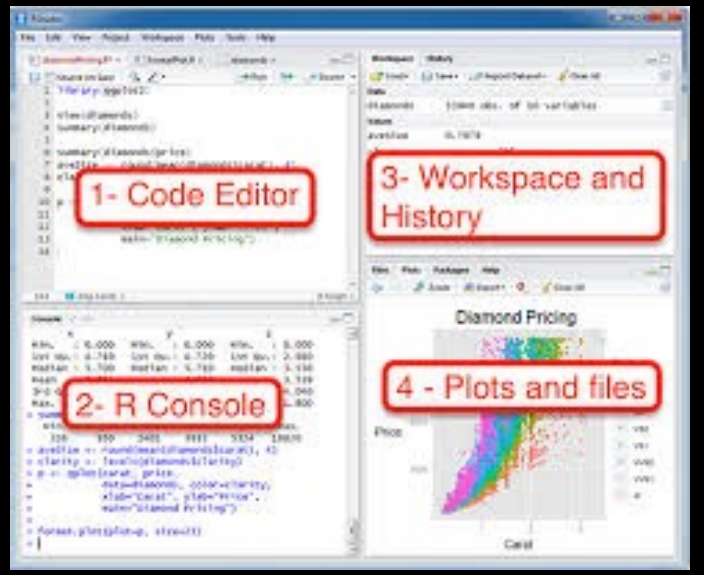
Working in R/R Studio

- **Always** use a text editor to save your work
  - Allows for repeatability when you save your code.
  - Allows you to add comments to scripts to remember what you have done.
  - Use # to make comments that won't be executed
  - Makes it easy to share code with collaborators

- When you type things into the console and execute them, **they are run but they are not saved**.

- To execute commands:

Mac: ⌘↵, PC: CTRL-R
Can highlight multiple lines of code and run at once



Exercise 1: a first session in R

- **Objective:** experiencing R/R studio

2. Enter and Import your data

Objective

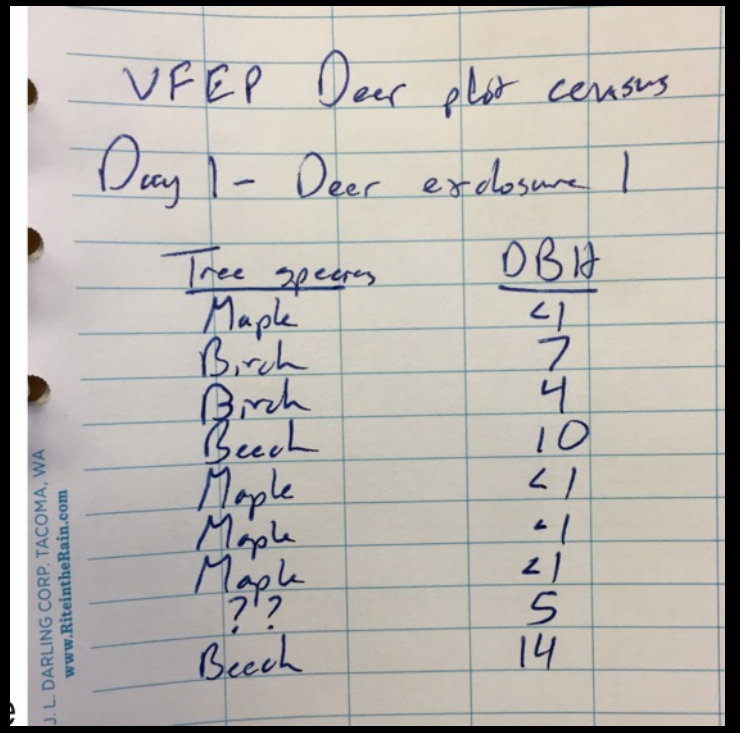
- To **teach the basic knowledge necessary to use R.**
    - **How to record your data?**
    - **How do you import them into R?**

    - **Experience R: live coding**

Record your data

- Most of the time have a data book where you write down your data, observations, etc.
- Most people use MS Excel to enter and store data from the notebook on the computer.
- **But… BEWARE of how data is recorded on excel**

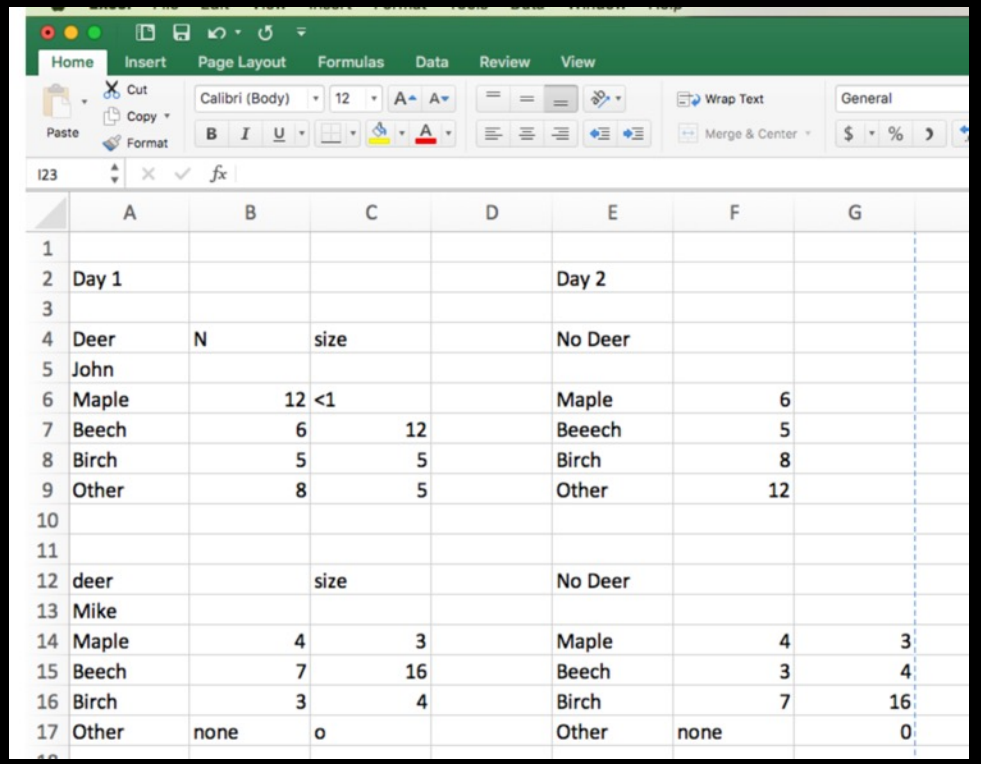
Hypothetical data on sizes of trees in deer exclosures



Record your data

- Most of the time have a data book where you write down your data, observations, etc.
- Most people use MS Excel to enter and store data from the notebook on the computer.
- **But… BEWARE of how data is recorded on excel**

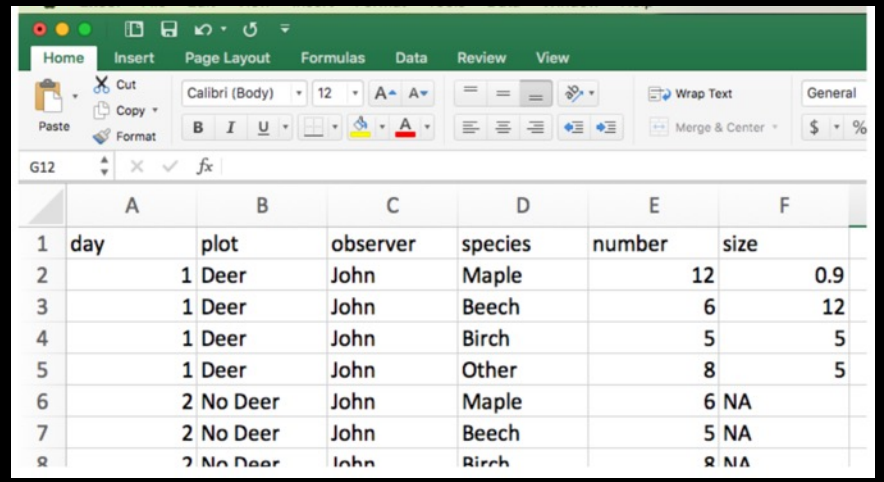
Hypothetical data on sizes of trees in deer exclosures



	A	B	C	D	E	F	G
1							
2	Day 1				Day 2		
3							
4	Deer	N	size		No Deer		
5	John						
6	Maple	12	<1		Maple	6	
7	Beech	6	12		Beeech	5	
8	Birch	5	5		Birch	8	
9	Other	8	5		Other	12	
10							
11							
12	deer		size		No Deer		
13	Mike						
14	Maple	4	3		Maple	4	3
15	Beech	7	16		Beech	3	4
16	Birch	3	4		Birch	7	16
17	Other	none	o		Other	none	0

Record your data: general rules

- Avoid spaces: use period "." or underscore "_".
- Keep column names short, simple and unique.
- **Be very careful of typos.**

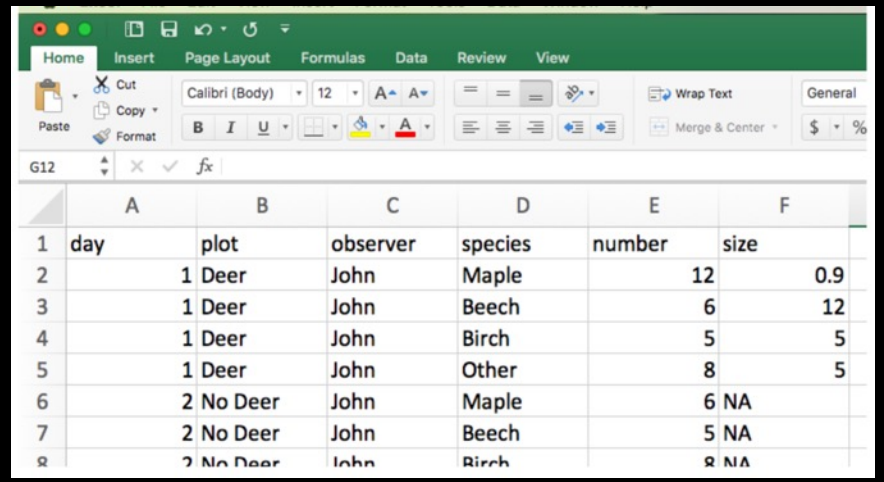


	A	B	C	D	E	F
1	day	plot	observer	species	number	size
2	1	Deer	John	Maple	12	0.9
3	1	Deer	John	Beech	6	12
4	1	Deer	John	Birch	5	5
5	1	Deer	John	Other	8	5
6	2	No Deer	John	Maple	6	NA
7	2	No Deer	John	Beech	5	NA
8	2	No Deer	John	Birch	8	NA

Record your data: general rules

- Avoid spaces: use period "." or underscore "_".
- Keep column names short, simple and unique.
- **Be very careful of typos.**
- **One variable per column (no merged column, no more than one).**
- Consistent unit throughout observations
- One observation per cell.

- Save as csv file



	A	B	C	D	E	F
1	day	plot	observer	species	number	size
2	1	Deer	John	Maple	12	0.9
3	1	Deer	John	Beech	6	12
4	1	Deer	John	Birch	5	5
5	1	Deer	John	Other	8	5
6	2	No Deer	John	Maple	6	NA
7	2	No Deer	John	Beech	5	NA
8	2	No Deer	John	Birch	8	NA

Import data in R