Course introduction

Michael Noonan

August 11, 2020

Biol 520C: Statistical modelling for biological data

Table of contents

1. Housekeeping
2. Design- vs. Model-based Inference
3. What is modelling?
4. Components of a model

Housekeeping

Slide title

Stuff

- Stuff

What this course is about

What this course is about

- Focus is on model-based inference

What this course is about

- Focus is on model-based inference (i.e., combining data with models to generate mechanistic descriptions of biological patterns)

What this course is about

- Focus is on model-based inference (i.e., combining data with models to generate mechanistic descriptions of biological patterns)

- Building from simple linear regression, you'll learn regression methods for handling the most routinely encountered features in biological data

What this course is about

- Focus is on model-based inference (i.e., combining data with models to generate mechanistic descriptions of biological patterns)

- Building from simple linear regression, you'll learn regression methods for handling the most routinely encountered features in biological data (hierarchical data structures, non-Gaussian error distributions, non-linearity, autocorrelation, etc...)

What this course is about

- Focus is on model-based inference (i.e., combining data with models to generate mechanistic descriptions of biological patterns)
- Building from simple linear regression, you'll learn regression methods for handling the most routinely encountered features in biological data (hierarchical data structures, non-Gaussian error distributions, non-linearity, autocorrelation, etc...)
- Emphasis on statistical best practices

What this course is about

- Focus is on model-based inference (i.e., combining data with models to generate mechanistic descriptions of biological patterns)

- Building from simple linear regression, you'll learn regression methods for handling the most routinely encountered features in biological data (hierarchical data structures, non-Gaussian error distributions, non-linearity, autocorrelation, etc...)

- Emphasis on statistical best practices

- How to use open source software (R) to apply these anaylses

What this course is not about

What this course is not about

- Basic statistics (concepts like means, medians, variances, probability distributions, regression should be familiar to you)

What this course is not about

- Basic statistics (concepts like means, medians, variances, probability distributions, regression should be familiar to you)

- Math

What this course is not about

- Basic statistics (concepts like means, medians, variances, probability distributions, regression should be familiar to you)
- Math... but you should be familiar with basic calculus (derivatives and integrals) and linear algebra (operations on vectors and matrices)

What this course is not about

- Basic statistics (concepts like means, medians, variances, probability distributions, regression should be familiar to you)

- Math... but you should be familiar with basic calculus (derivatives and integrals) and linear algebra (operations on vectors and matrices)

- Computer programming (we will be using `R`, but the course is not focused on high level coding)

What this course is not about

- Basic statistics (concepts like means, medians, variances, probability distributions, regression should be familiar to you)
- Math... but you should be familiar with basic calculus (derivatives and integrals) and linear algebra (operations on vectors and matrices)
- Computer programming (we will be using `R`, but the course is not focused on high level coding)
- Methods for handling *ad hoc*, corner cases

Design- vs. Model-based Inference

The Scientific Method

The Scientific Method

Science is a process of learning about nature. As scientists, we weigh competing ideas about how the world works (hypotheses) against observations (data).

The Scientific Method

Science is a process of learning about nature. As scientists, we weigh competing ideas about how the world works (hypotheses) against observations (data). But our descriptions of the world are almost always incomplete, our observations have error, and important data is often missing...

The Scientific Method

Science is a process of learning about nature. As scientists, we weigh competing ideas about how the world works (hypotheses) against observations (data). But our descriptions of the world are almost always incomplete, our observations have error, and important data is often missing... So, how do we accurately compare what we observe with what we hypothesize without bias?

The Scientific Method

Science is a process of learning about nature. As scientists, we weigh competing ideas about how the world works (hypotheses) against observations (data). But our descriptions of the world are almost always incomplete, our observations have error, and important data is often missing... So, how do we accurately compare what we observe with what we hypothesize without bias?

Statistics

The Scientific Method

Science is a process of learning about nature. As scientists, we weigh competing ideas about how the world works (hypotheses) against observations (data). But our descriptions of the world are almost always incomplete, our observations have error, and important data is often missing... So, how do we accurately compare what we observe with what we hypothesize without bias?

Statistics

The process of making scientific inference can be split into two broad categories:

The Scientific Method

Science is a process of learning about nature. As scientists, we weigh competing ideas about how the world works (hypotheses) against observations (data). But our descriptions of the world are almost always incomplete, our observations have error, and important data is often missing... So, how do we accurately compare what we observe with what we hypothesize without bias?

Statistics

The process of making scientific inference can be split into two broad categories: design-based

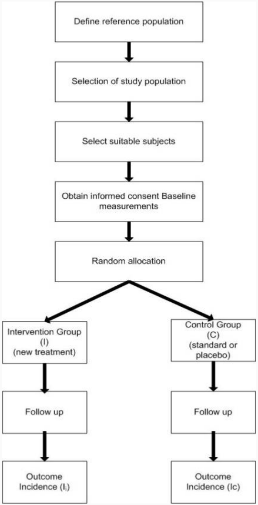
The Scientific Method

Science is a process of learning about nature. As scientists, we weigh competing ideas about how the world works (hypotheses) against observations (data). But our descriptions of the world are almost always incomplete, our observations have error, and important data is often missing... So, how do we accurately compare what we observe with what we hypothesize without bias?

Statistics

The process of making scientific inference can be split into two broad categories: design-based and model-based

Design-based inference


Define reference population

Selection of study population

Select suitable subjects

Obtain informed consent Baseline measurements

Random allocation

Intervention Group (I) (new treatment)

Control Group (C) (standard or placebo)

Follow up

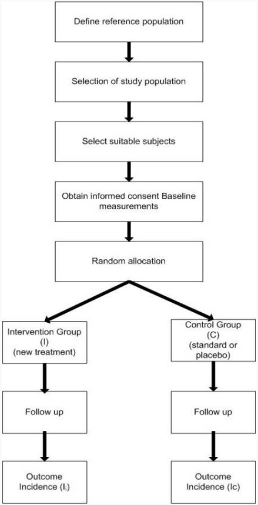
Follow up

Outcome Incidence (I_I)

Outcome Incidence (Ic)

Design-based inference

- In design-based inference, most of the focus is on experimental design



Define reference population

Selection of study population

Select suitable subjects

Obtain informed consent Baseline measurements

Random allocation

Intervention Group (I) (new treatment)

Control Group (C) (standard or placebo)

Follow up

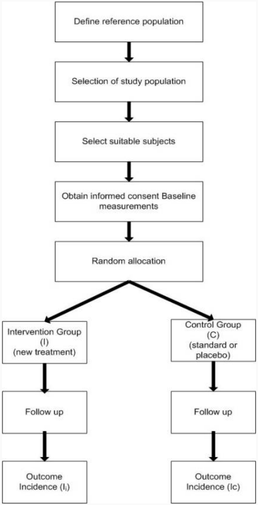
Follow up

Outcome Incidence (I_I)

Outcome Incidence (Ic)

Design-based inference

- In design-based inference, most of the focus is on experimental design
- You assume your sample is a random sample of a target population



Define reference population

Selection of study population

Select suitable subjects

Obtain informed consent Baseline measurements

Random allocation

Intervention Group (I) (new treatment)

Control Group (C) (standard or placebo)

Follow up

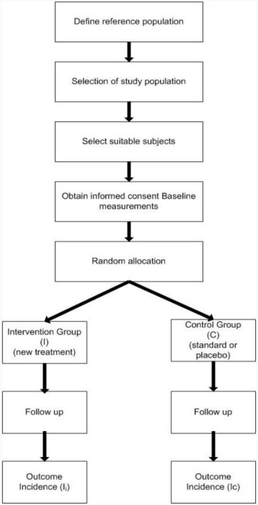
Follow up

Outcome Incidence (I_I)

Outcome Incidence (Ic)

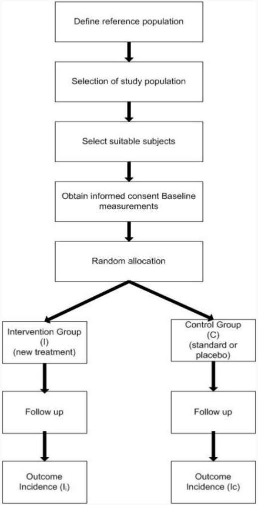
Design-based inference

- In design-based inference, most of the focus is on experimental design
- You assume your sample is a random sample of a target population
- Inference is focused towards a target population


Define reference population
Selection of study population
Select suitable subjects
Obtain informed consent Baseline measurements
Random allocation
Intervention Group (I) (new treatment)
Control Group (C) (standard or placebo)
Follow up
Follow up
Outcome Incidence (I_I)
Outcome Incidence (Ic)

Design-based inference

- In design-based inference, most of the focus is on experimental design
- You assume your sample is a random sample of a target population
- Inference is focused towards a target population
- Goal is to be able to demonstrate that ‘x causes y’



Define reference population

Selection of study population

Select suitable subjects

Obtain informed consent Baseline measurements

Random allocation

Intervention Group (I) (new treatment)

Control Group (C) (standard or placebo)

Follow up

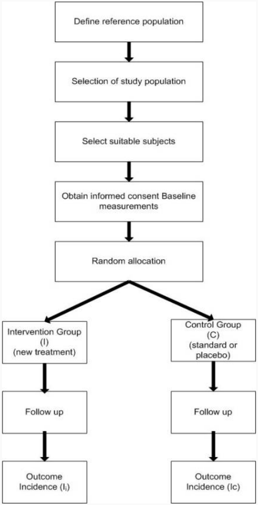
Follow up

Outcome Incidence (I_i)

Outcome Incidence (Ic)

Design-based inference

- In design-based inference, most of the focus is on experimental design
- You assume your sample is a random sample of a target population
- Inference is focused towards a target population
- Goal is to be able to demonstrate that 'x causes y'
- Data are typically analysed by comparing means and variances across groups (e.g., ANOVAs, *t*-tests, etc...)



Define reference population

Selection of study population

Select suitable subjects

Obtain informed consent Baseline measurements

Random allocation

Intervention Group (I) (new treatment)

Control Group (C) (standard or placebo)

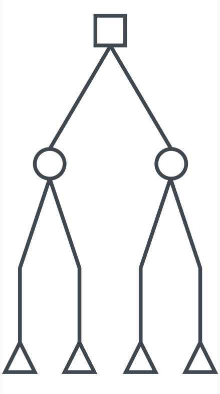
Follow up

Follow up

Outcome Incidence (I_I)

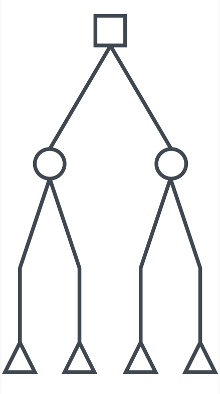
Outcome Incidence (Ic)

Worklow of design-based inference



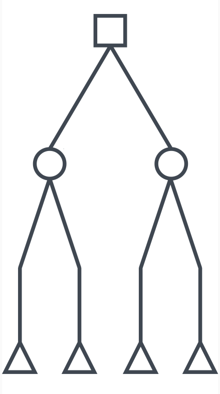
Worklow of design-based inference

- Devise a hypotheses



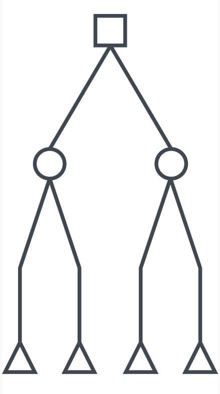
Worklow of design-based inference

- Devise a hypotheses
- Devise an experiment with outcomes that will clearly accept or reject the hypothesis



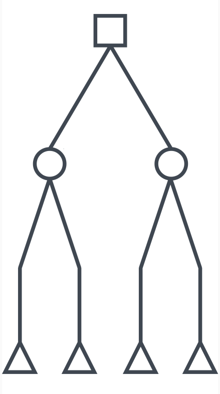
Worklow of design-based inference

- Devise a hypotheses
- Devise an experiment with outcomes that will clearly accept or reject the hypothesis
- Carry out the experiment so as to get a clean result



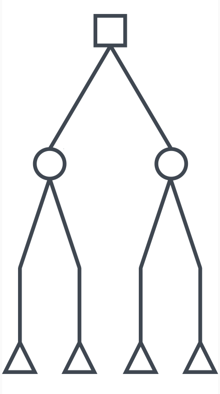
Worklow of design-based inference

- Devise a hypotheses
- Devise an experiment with outcomes that will clearly accept or reject the hypothesis
- Carry out the experiment so as to get a clean result
- Recycle the procedure to refine the remaining possibilities (Platt, 1964)



Worklow of design-based inference

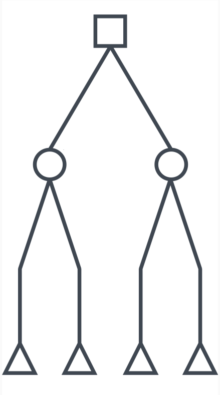
- Devise a hypotheses
- Devise an experiment with outcomes that will clearly accept or reject the hypothesis
- Carry out the experiment so as to get a clean result
- Recycle the procedure to refine the remaining possibilities (Platt, 1964)



Core of design-based inference is confronting single hypotheses with data

Limitations of design-based inference

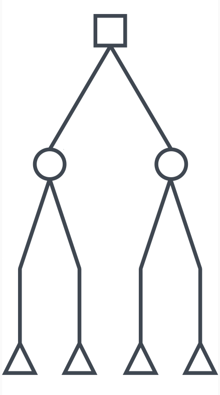
Platt's decision tree is based on:



Limitations of design-based inference

Platt's decision tree is based on:

i) Clear, distinct hypotheses

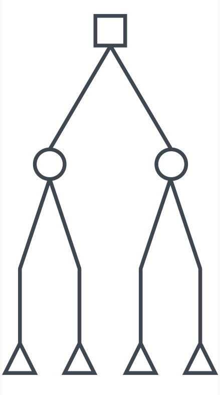


Limitations of design-based inference

Platt's decision tree is based on:

i) Clear, distinct hypotheses

ii) Unambiguous outcomes



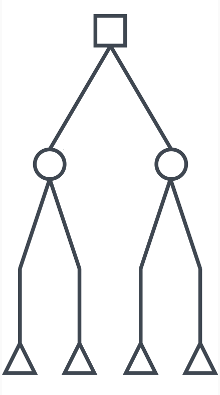
Limitations of design-based inference

Platt's decision tree is based on:

i) Clear, distinct hypotheses

ii) Unambiguous outcomes

iii) A relationship between statistical significance and biological relevance



Limitations of design-based inference

Many biological processes have long time-scales.

Long time-scales

Long time-scales

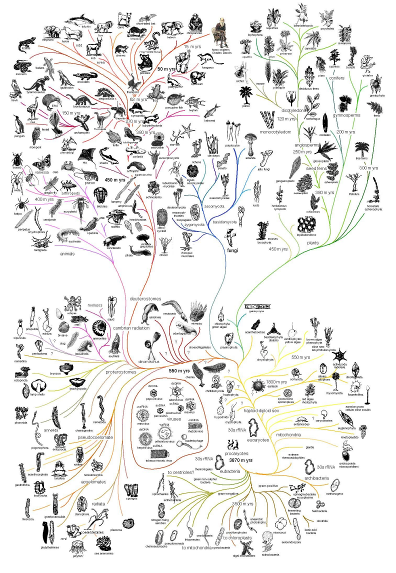
Sperm whales (*Physeter macrocephalus*) can live for 70+ years

Long time-scales

Sperm whales (*Physeter macrocephalus*) can live for 70+ years

Bristlecone pines (*Pinus longaeva*) live for thousands of years

Long time-scales



Limitations of design-based inference

Many biological processes have long time-scales.

Limitations of design-based inference

Many biological processes have long time-scales.

Many biological systems have very poor reproducibility.

Poor reproducibility

Limitations of design-based inference

Many biological processes have long time-scales.

Many biological systems have very poor reproducibility.

How can you design a controlled experiment in a wild population?

Limitations of design-based inference

What if you're interested in species conservation?

Limitations of design-based inference

What if you're interested in species conservation?

In 1988 the wild pop. of black footed ferrets (*Mustela nigripes*) was down to 18 ind.

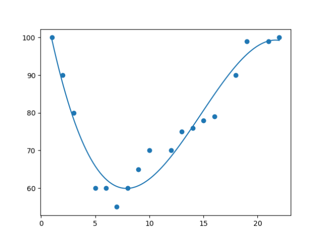
Limitations of design-based inference

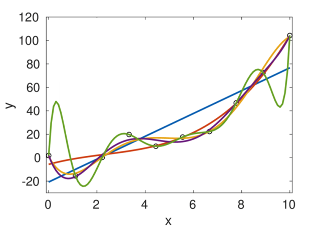
What if you're interested in species conservation?

In 1988 the wild pop. of black footed ferrets (*Mustela nigripes*) was down to 18 ind.

What do you do if a power analysis says you need 20 animals?

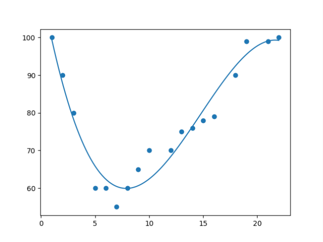
Model-based inference

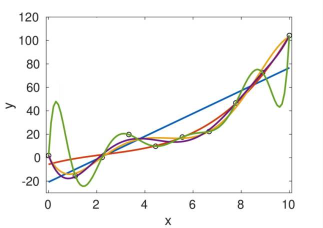




Model-based inference

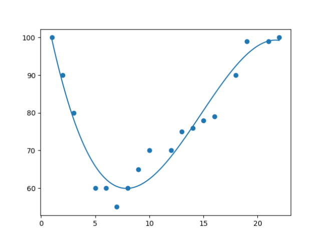
- In model-based inference, most of the focus is on identifying an unknown deterministic model

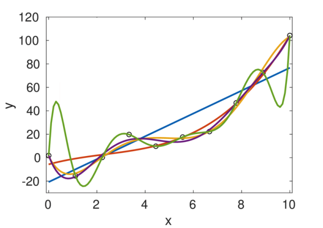




Model-based inference

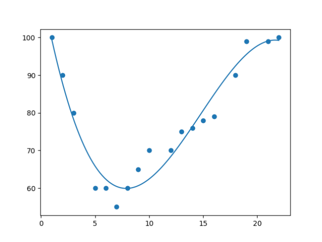
- In model-based inference, most of the focus is on identifying an unknown deterministic model
- Inference can be extrapolated beyond the target population

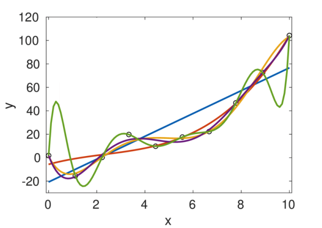




Model-based inference

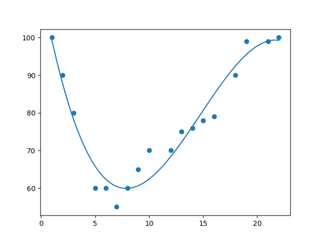
- In model-based inference, most of the focus is on identifying an unknown deterministic model
- Inference can be extrapolated beyond the target population
- Goal is to provide a theoretical framework for why ‘x causes y’

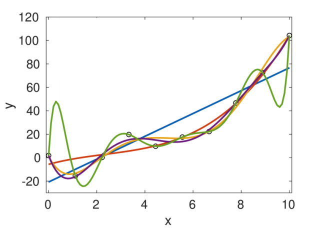




Model-based inference

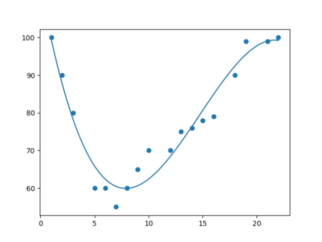
- In model-based inference, most of the focus is on identifying an unknown deterministic model
- Inference can be extrapolated beyond the target population
- Goal is to provide a theoretical framework for why ‘x causes y’
- In model-based inference you make distributional assumptions to make your response a random variable

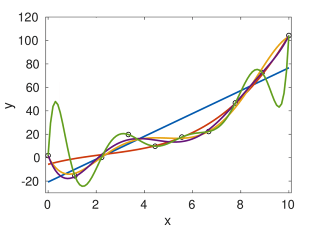




Model-based inference

- In model-based inference, most of the focus is on identifying an unknown deterministic model
- Inference can be extrapolated beyond the target population
- Goal is to provide a theoretical framework for why ‘x causes y’
- In model-based inference you make distributional assumptions to make your response a random variable
- Data are typically analysed by fitting a model to data and interpreting the parameter estimates





What is modelling?

Hypothesis vs. model

Hypothesis vs. model

Hypothesis:

Hypothesis vs. model

Hypothesis: An idea, supposition, or otherwise unproven theory used as the basis for further investigation.

Hypothesis vs. model

Hypothesis: An idea, supposition, or otherwise unproven theory used as the basis for further investigation.

Model:

Hypothesis vs. model

Hypothesis: An idea, supposition, or otherwise unproven theory used as the basis for further investigation.

Model: A generalised description of some phenomenon.

Hypothesis vs. model

Hypothesis: An idea, supposition, or otherwise unproven theory used as the basis for further investigation.

Model: A generalised description of some phenomenon.

$$\text{Model} \neq \text{Hypothesis}$$

One hypothesis, many models

One hypothesis, many models

A single hypothesis can be represented by multiple models.

One hypothesis, many models

A single hypothesis can be represented by multiple models.

Hypothesis: Body mass M increases with age L

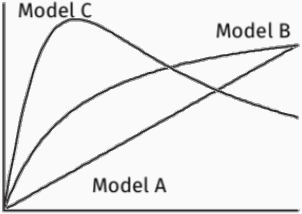
One hypothesis, many models

A single hypothesis can be represented by multiple models.

Hypothesis: Body mass M increases with age L

Models:

- $M = aL$ Model A: Body mass is proportional to age
- $M = \frac{AL}{1+bL}$ Model B: Body mass saturates as age increases
- $M = aLe^{-bL}$ Model C: Body mass increases and then decreases as age increases



General points about modelling

THE UNIVERSITY OF BRITISH COLUMBIA
Okanagan Campus

General points about modelling

The equation of a model is a very specific expression of the hypothesis.

General points about modelling

The equation of a model is a very specific expression of the hypothesis. In other words, models help us clarify verbal descriptions of nature and mechanisms.

General points about modelling

The equation of a model is a very specific expression of the hypothesis. In other words, models help us clarify verbal descriptions of nature and mechanisms.

Models help us understand which parameters and processes are important, and which ones are not.

General points about modelling

The equation of a model is a very specific expression of the hypothesis. In other words, models help us clarify verbal descriptions of nature and mechanisms.

Models help us understand which parameters and processes are important, and which ones are not.

No model is completely correct.

No model is completely correct

No model is completely correct

The goal of modelling is not to provide a perfect description of the world

No model is completely correct

The goal of modelling is not to provide a perfect description of the world, but to distill a process down to the most important components

No model is completely correct

The goal of modelling is not to provide a perfect description of the world, but to distill a process down to the most important components

Complicated models with lots of parameters usually provide better fits

No model is completely correct

The goal of modelling is not to provide a perfect description of the world, but to distill a process down to the most important components

Complicated models with lots of parameters usually provide better fits, but if the model is as complicated as nature itself why bother with modelling? Just go for a walk in the woods and be happy.

No model is completely correct

The goal of modelling is not to provide a perfect description of the world, but to distill a process down to the most important components

Complicated models with lots of parameters usually provide better fits, but if the model is as complicated as nature itself why bother with modelling? Just go for a walk in the woods and be happy.

Don’t fall in love with a model, the important thing is the system

Model complexity

Model complexity

Complex models provide more numerical precision

Model complexity

Complex models provide more numerical precision, but simple models are more interpretable.

Model complexity

Complex models provide more numerical precision, but simple models are more interpretable.

Simple models risk leaving out important parameters

Model complexity

Complex models provide more numerical precision, but simple models are more interpretable.

Simple models risk leaving out important parameters, complex models need lots of data for good parameter estimation

Model complexity

Complex models provide more numerical precision, but simple models are more interpretable.

Simple models risk leaving out important parameters, complex models need lots of data for good parameter estimation

How complex should a model be?

Model complexity

Complex models provide more numerical precision, but simple models are more interpretable.

Simple models risk leaving out important parameters, complex models need lots of data for good parameter estimation

How complex should a model be?

Short answer: Let the data tell you.

Model complexity

Complex models provide more numerical precision, but simple models are more interpretable.

Simple models risk leaving out important parameters, complex models need lots of data for good parameter estimation

How complex should a model be?

Short answer: Let the data tell you.

Long answer: There are methods for this that we'll cover in later lectures.

Components of a model

Model Components

Model Components

Models are comprised of two main components:

Model Components

Models are comprised of two main components:

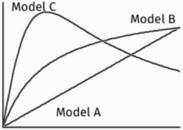
Deterministic part: Describes the shape of the relationship (i.e., your hypothesis).

Model Components

Models are comprised of two main components:

Deterministic part: Describes the shape of the relationship (i.e., your hypothesis).

- Model A: $M = aL$
- Model B: $M = \frac{AL}{1+bL}$
- Model C: $M = aLe^{-bL}$

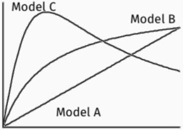


Model Components

Models are comprised of two main components:

Deterministic part: Describes the shape of the relationship (i.e., your hypothesis).

- Model A: $M = aL$
- Model B: $M = \frac{AL}{1+bL}$
- Model C: $M = aLe^{-bL}$



Stochastic part: Describes the randomness of the process (i.e., captures the noise in a system).

Deterministic vs. Stochastic

Deterministic vs. Stochastic

Deterministic models

Deterministic vs. Stochastic

Deterministic models

- No components are uncertain

Deterministic vs. Stochastic

Deterministic models

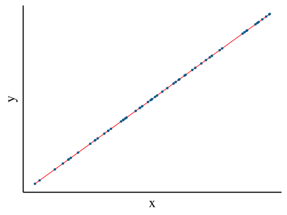
- No components are uncertain

- Outcome is always the same

Deterministic vs. Stochastic

Deterministic models

- No components are uncertain
- Outcome is always the same
- $y_i = \beta_0 + \beta_1 x_i$



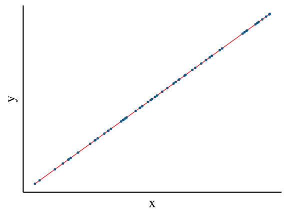
Deterministic vs. Stochastic

Deterministic models

- No components are uncertain

- Outcome is always the same

- $y_i = \beta_0 + \beta_1 x_i$

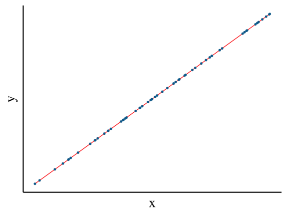


Stochastic models

Deterministic vs. Stochastic

Deterministic models

- No components are uncertain
- Outcome is always the same
- $y_i = \beta_0 + \beta_1 x_i$



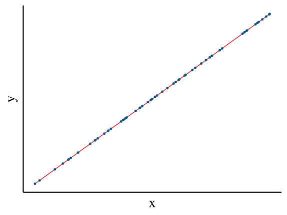
Stochastic models

- Some components are uncertain and characterised by probability distributions

Deterministic vs. Stochastic

Deterministic models

- No components are uncertain
- Outcome is always the same
- $y_i = \beta_0 + \beta_1 x_i$



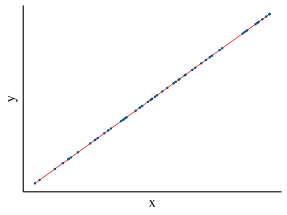
Stochastic models

- Some components are uncertain and characterised by probability distributions
- Outcome is variable

Deterministic vs. Stochastic

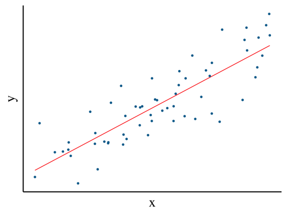
Deterministic models

- No components are uncertain
- Outcome is always the same
- $y_i = \beta_0 + \beta_1 x_i$



Stochastic models

- Some components are uncertain and characterised by probability distributions
- Outcome is variable
- $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$



References

Platt, J.R. (1964). Strong inference. *science*, 146, 347–353.