Lecture 2: Statistical modelling

STA238: Probability, Statistics, and Data Analysis II

Fred Song

Week of July 8th, 2024 Lec 2

What is a statistical model?

> *Something which accurately resembles or represents something else, esp. on a small scale; a person or thing that is the likeness of another.*

> *A simplified or idealized description or conception of a particular system, situation, or process, often in mathematical terms, that is put forward as a basis for theoretical or empirical understanding, or for calculations, predictions, etc.; a conceptual or mental representation of something.*

From Oxford English Dictionary, https://www.oed.com/dictionary/model_n

Example: Newton's second law of motion

$$F = ma$$

- It is an idealized description of one aspect of physical motions.
- It provides a theoretical understanding on the relationship between force, mass, and acceleration.

> It's an example of a *deterministic* model. There is no uncertainty involved in the model.

A statistical model

> A **statistical model** is an idealization of a data-generating process.

- Specifically, we model data-generating processes using probability distributions.
- The distributions describe the uncertainty involved in the data-generating process.

  e.g., measurement errors, unknown variables effecting the observations, subjective judgement, inherent randomness, ...

Is a model true?

> *All models are wrong, but some are useful. – George Box.*

Reality

Observations

Information

Real & complex

Messy

Data

Theory

Explanations

Patterns

Ideal & simplified

Abstract

Model

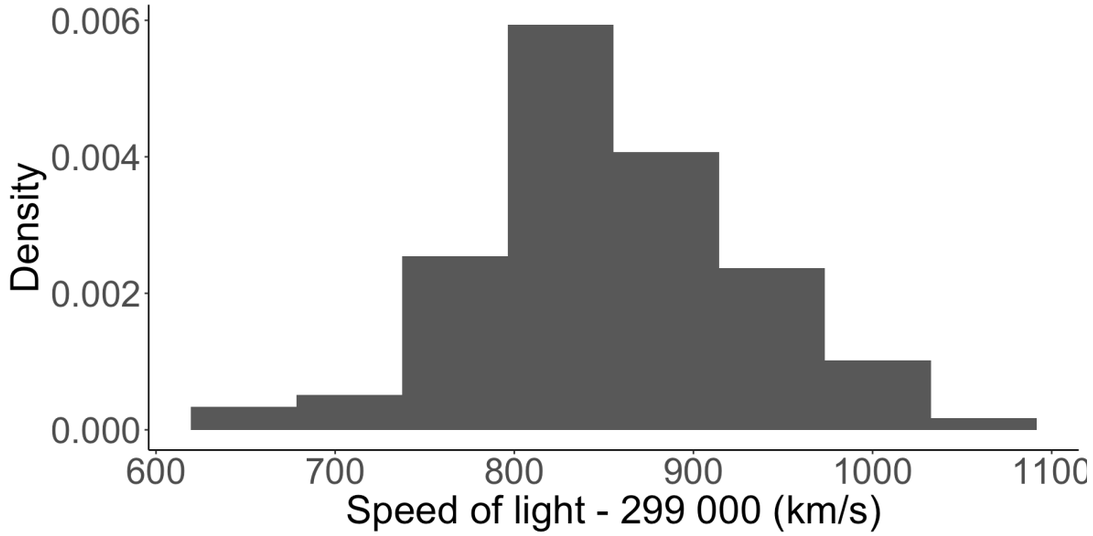
Example: The speed of light

> In km/s minus 299 000.
>
> See Section 1.6 and 17.1 of Dekking *et al.* for details.
>
> They are available as `morley` data set in R.

Speed of light by Albert Michelson, 1879.

850	1000	960	830	880	880	890	910	890	870
740	980	940	790	880	910	810	920	840	870
900	930	960	810	880	850	810	890	780	810
1070	650	940	880	860	870	820	860	810	740
930	760	880	880	720	840	800	880	760	810
850	810	800	830	720	840	770	720	810	940
950	1000	850	800	620	850	760	840	790	950
980	1000	880	790	860	840	740	850	810	800
980	960	900	760	970	840	750	850	820	810
880	960	840	800	950	840	760	780	850	870

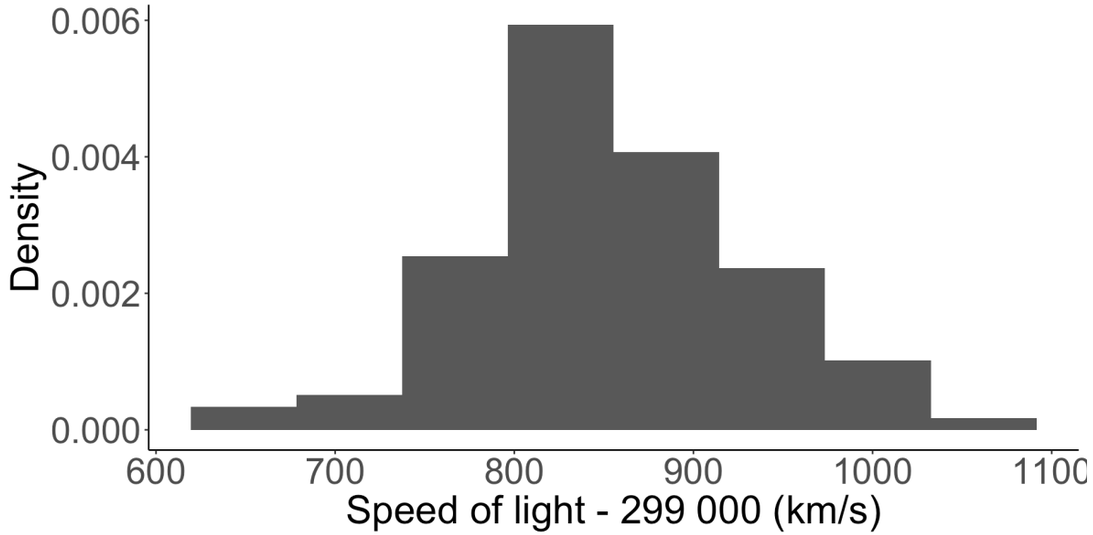
Histogram of speed of light



- Histograms is an approximation of a probability density function.
- Which distribution is the histogram approximating?

x_1	$X_1 \sim F_1$
x_2	$X_2 \sim F_2$
x_3	$X_3 \sim F_3$
$\vdots$	$\vdots$

Histogram of speed of light



- We are interested in learning about a single quantity — the speed of light.
- We assume all n data values are *repeated* realizations from a *single* distribution.
- We treat these data as realizations of a **random sample**.

Random sample

- In other words,

$$X_1, X_2, \ldots, X_n \overset{i.i.d}{\sim} F$$

> i.i.d: independent and identically distributed.
>
> $X \sim F$: X follows a distribution defined by the cumulative distribution function F.

> A **random sample** is a collection of random variables X_1, X_2, ..., X_n that have the same probability distribution and are mutually independent.

Exercises

Suppose X_1, X_2, ..., X_n are a random sample from a distribution with mean 5 and variance 2. What is the mean and variance of their sum?

Is the speed of light data set a random sample?

Do they follow the same distribution?

Are the observations independent?

Is the speed of light data set a random sample?

Do they follow the same distribution?

- The measurements were collected under the same experimental conditions.
- They were estimates for a one “true” value.

Are the observations independent?

- Physical independence – the events associated with experiments have no physical connections – implies stochastic independence.
- When collecting the speed of light data, Michelson ensured the measurements had no influence to each other.

Statistical model for repeated measurements (univariate)

- The probability distribution F is the **statistical model**.
- We assume that there exists one *true distribution*.[†]

[†]At least until we talk about Bayesian estimation.

A data set consisting of repeated measurements x_1, x_2, ..., x_n of the same quantity is modelled as the realization of a random sample X_1, X_2, ..., X_n.

The model may include a partial specification of the probability distribution of each X_i.

Example: A statistical model for the speed of light measurements

What we know/believe ...

- The speed of light in a vacuum is constant.
- The observed randomness is due to measurement errors.
- The experiment was carefully designed to avoid any systematic error.

How we may model our knowledge/belief ...

- The observed values are centred around the speed of light in a vacuum.
- The expectation of the errors is 0.
- The errors have a finite variance.

> We can model $\text{Measurement} = \text{True speed} + \text{Error}$, which implies that $\mathbb{E}(\text{Measurement}) = \text{True speed}$ and $\text{Var}(\text{Measurement}) = \text{Var}(\text{Error}) < \infty$.

Example: Software reliability data

Time between successive failures in CPU seconds by Musa et al, 1987.

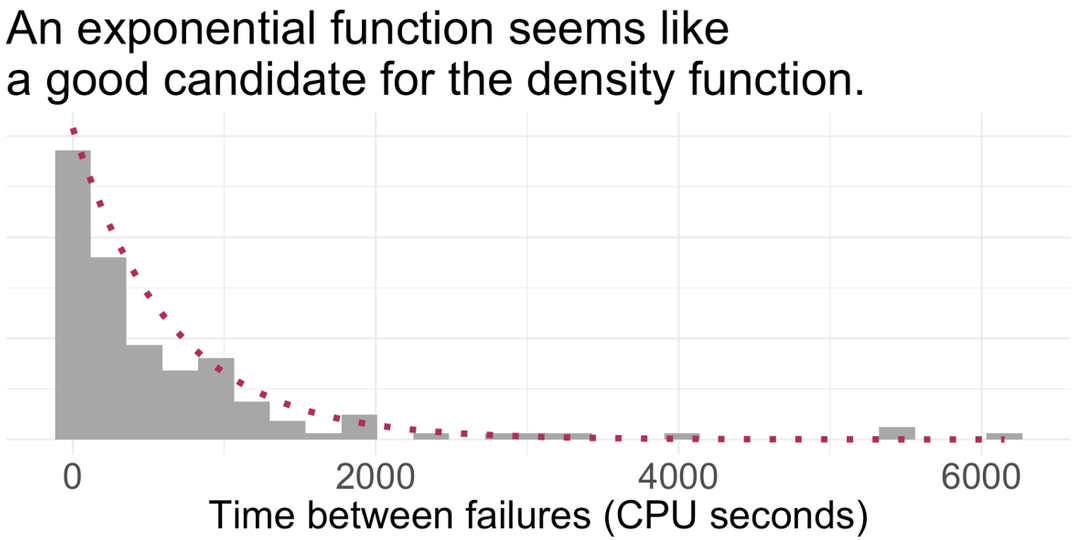
30	26	65	233	379	983	0	75
113	114	176	134	44	707	3110	4
81	325	58	357	129	33	1247	82
115	55	457	193	810	868	943	5509
9	242	300	236	290	724	700	100
2	68	97	31	300	2323	875	10
91	422	263	369	529	2930	245	1071
112	180	452	748	281	1461	729	371
15	10	255	0	160	843	1897	790
138	1146	197	232	828	12	447	6150
50	600	193	330	1011	261	386	3321
77	15	6	365	445	1800	446	1045
24	36	79	1222	296	865	122	648
108	4	816	543	1755	1435	990	5485
88	0	1351	10	1064	30	948	1160
670	8	148	16	1783	143	1082	1864
120	227	21	529	860	108	22	4116

See Section 15.3 of Dekking *et al.* for details.

For each failure in a certain software, the length of the interval from the previous, or the start of the experiment, is recorded. What is a reasonable model?

- Times are nonnegative.
- We may assume software failures are a Poisson process.
  - Failures are mutually independent.
  - Failures can occur at any moment with the same likelihood.

Example: Software reliability data



- Based on the assumptions, we may partially specify the model.
- Specifying the **model parameter(s)**, θ fully specifies the model.
  - In the case of $\text{Exp}(\lambda)$, $\theta = \lambda$.
- The partially specified model is called a **parametric model**.

$$\{F_\theta : \theta \in \Theta\}$$

- The *true* parameter(s) specifies the *true* distribution.

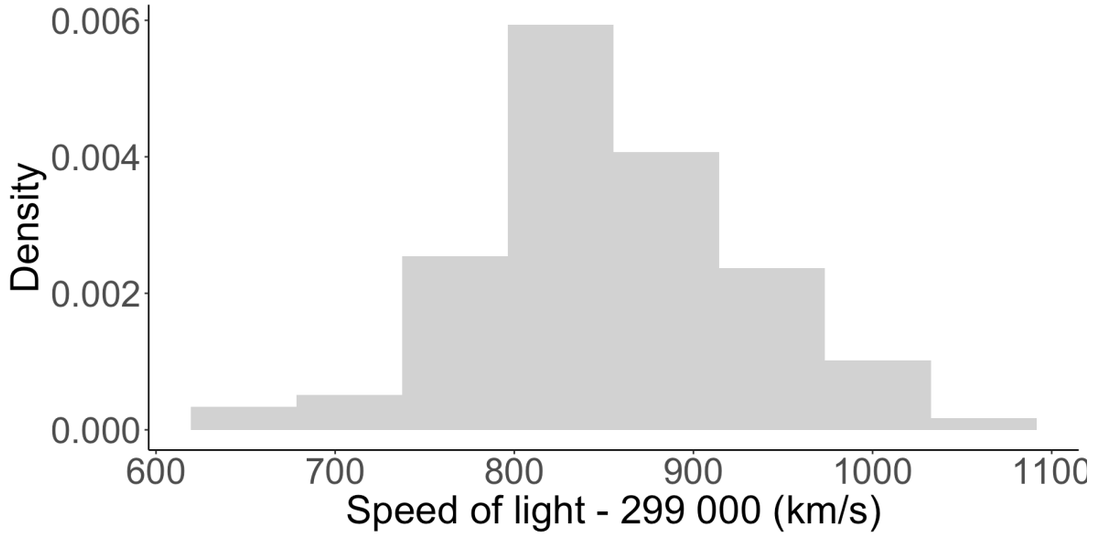
Exercises

What is a reasonable candidate distribution to model each of the following data? Assume they are random samples. What simplifying assumptions are you making about the context of the data?

- The lifetime of a battery under continuous use.
- The number of wins by the Leafs out of 82 regular season games.
- Average class grades of *Advanced Functions* from Ontario high schools.

Estimating features of the model distribution

Example: The true speed of light



Based on the model, how can we estimate the *true* speed of light in a vacuum?

$$\begin{aligned}&\mathbb{E}(\text{Measurement})\\&= \mathbb{E}(\text{True speed}) + \mathbb{E}(\text{Error})\\&= \text{True speed} + 0\end{aligned}$$

- Estimate the mean of the model distribution.
- We already know how we can estimate the mean.

Recall: Chebyshev's Inequality

Consider a random variable Y with $\mathbb{E}(Y) < \infty$ and $\mathbf{Var}(Y) < \infty$, and a constant $\varepsilon > 0$. Then

$$P(|Y - \mathbb{E}(Y)| \geq \varepsilon) \leq \frac{1}{\varepsilon^2}\mathbf{Var}(Y).$$

Proof:

Recall: The Weak Law of Large Numbers

Consider a sequence of independent random variables X_1, X_2, ..., with each having the same mean $\mu < \infty$ and the same variance $\sigma^2 < \infty$. Then for all $\varepsilon > 0$,

$$\lim_{n \to \infty} P(|\bar{X}_n - \mu| \geq \varepsilon) = 0,$$

where $\bar{X}_n = \sum_{i=1}^{n} X_i \big/ n$. In other words,

$$\bar{X}_n \xrightarrow{p} \mu \quad \text{as} \quad n \to \infty.$$

Proof:

Recall: The Weak Law of Large Numbers

Consider a sequence of independent random variables X_1, X_2, ..., with each having the same mean $\mu < \infty$ and the same variance $\sigma^2 < \infty$. Then for all $\varepsilon > 0$,

$$\lim_{n \to \infty} P(|\bar{X}_n - \mu| \geq \varepsilon) = 0,$$

where $\bar{X}_n = \sum_{i=1}^{n} X_i \big/ n$. In other words,

$$\bar{X}_n \xrightarrow{p} \mu \quad \text{as} \quad n \to \infty.$$

- A random sample is a sequence of independent and identically distributed random variables.
- The sample mean of a random sample estimates the expectation of the model distribution.
- The sample mean is an example of a **sample statistic**.

Sample statistics

- With a model adequately represents the data, sample statistics reflect features of the distribution.
- Note that a sample statistic is a random variable.

> Consider a random sample X_1, X_2, ..., X_n. A **sample statistic** is a function of the random sample,
>
> $$h(X_1, X_2, \ldots, X_n).$$

Example: Variance of speed of light measurements

The speed of light in a vacuum is now defined as exactly 299 792.458 km/s.

Let $X_i \overset{i.i.d}{\sim} F_\theta$ for $i = 1, \ldots, n$ represent speed of light measurements in km/s where $\theta = (\mu, \sigma^2)$. $\mu = 299\,792.458$ is the mean and σ^2 is the unknown variance of the measurements.

Suppose we estimate the true variance σ^2 with

$$\tilde{s}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2.$$

Show that $\tilde{s}_n^2 \xrightarrow{p} \sigma^2$. What extra conditions are required on X_i?

X_i's are i.i.d random variables.

$\implies (X_i - \mu)^2$ are i.i.d random variables.

$\implies \tilde{s}_n^2$ is a average of i.i.d random variables.

Required conditions for the LLN:

1. $\mathbb{E}\left[(X_i - \mu)^2\right] = \sigma^2 < \infty$
2. $\text{Var}\left[(X_i - \mu)^2\right] < \infty$

Then, $\tilde{s}_n^2$ converges in probability to its mean.

$$\mathbb{E}(\tilde{s}_n^2) = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2\right) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i - \mu)^2$$
$$= \frac{1}{n} \cdot n \cdot \sigma^2 = \sigma^2$$
$$\implies \tilde{s}_n^2 \xrightarrow{p} \sigma^2$$

Exercises

(See Section 17.2 from Dekking et al.)

Let Y_1, Y_2, ..., Y_n be a random sample from a cumulative distribution function F and let F_n be the empirical distribution function.

$$F_n(a) = \frac{\text{number of } Y_i \text{ in } (-\infty, a]}{n}$$

Let $G_a = F(a)$ and $G_{n,a} = F_n(a)$ for some constant a. Show that $G_{n,a} \xrightarrow{p} G_a$.

> To estimate the full distribution, you need to repeat the process for every possible value of a.

Exercises

Toronto Maple Leafs won 50 games out of 82 regular season games during 2022-23 NHL season.

Assuming the number of wins in the 2023-24 season follows the binomial distribution as the number of wins in the 2022-23 season, what is the probability that the Leafs win more than 60 games in 2023-24?

What assumptions are easily violated?

- This is an example of a parametric model — however unrealistic.
- Estimating the parameter(s) fully estimates the parametric models.

Sample statistics and distribution features

(Table 17.2 from Dekking et al.)

Sample statistic	Distribution feature

Graphical	*Functions*
Empirical cumulative distribution function	Cumulative distribution function
Kernel density estimate and histogram	Probability density function and mass function
Numerical	*Parameters*
Sample mean	Expectation
Sample median	Median
Empirical quantile	Quantile
Sample variance	Variance
Sample standard deviation	Standard deviation
Median absolute deviation	$F^{\text{inv}}(0.75) - F^{\text{inv}}(0.5)$, for symmetric F[‡]

[‡]See Section 17.2 from Dekking et al.

Summary

- A **statistical model** is a simplified description of a data-generating process using a probability distribution.
- We model repeated univariate measurements as independent and identically distributed random variables when the assumptions are justified.
- Knowledge about the process can help us partially specify the data-generating process as a parametric model.
- The law of large numbers allows us to estimate features of model distributions using sample statistics.

Parameters and statistics

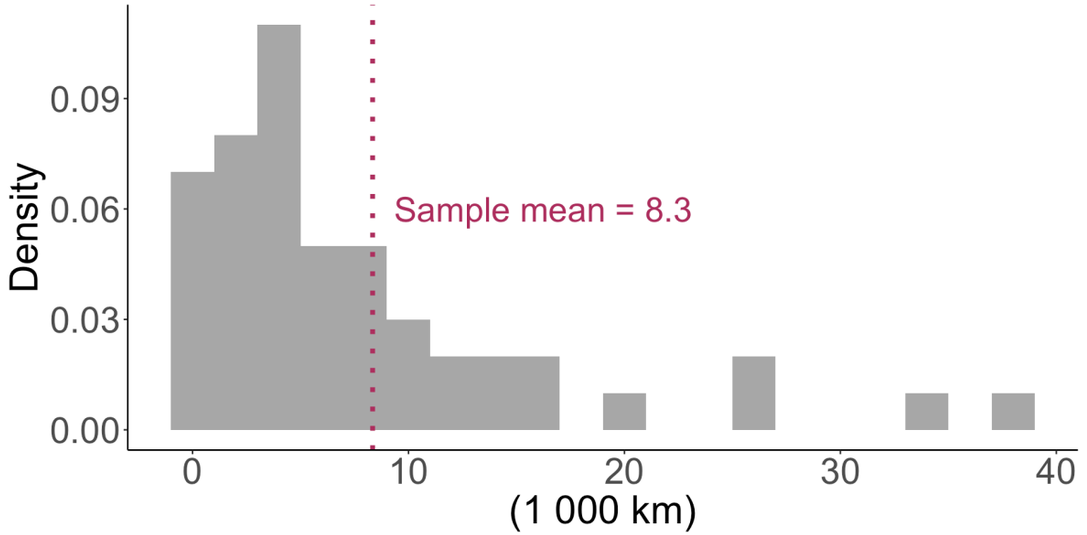
Example: Streetcar failure time

Recall the streetcar failure data example.

Assuming the data values are a random sample, we can model them as realizations of $X \sim$_____.

The model parameter is _____.

We may approximate the parameter as _____ based on the sample mean.



The data values aren't real values.

Exercises

Consider the following models. What does the sample mean tell us about the model parameter(s)?

- $\text{Binom}(82, p)$ to model the number of wins by the Leafs out of 82 regular season games.
- $\mathcal{N}(\mu, \sigma^2)$ to model the average class grades of *Advanced Functions* from Ontario high schools.

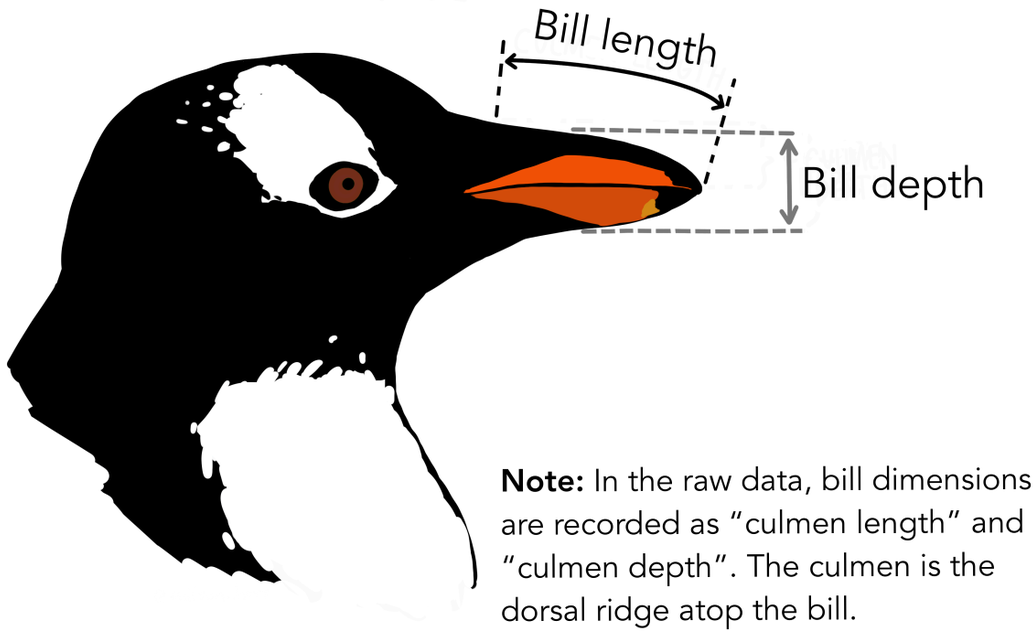
Exercises

Consider the following models. What does the sample mean tell us about the model parameter(s)?

- $\text{Binom}(82, p)$ to model the number of wins by the Leafs out of 82 regular season games.
- $\mathcal{N}(\mu, \sigma^2)$ to model the average class grades of *Advanced Functions* from Ontario high schools.

> Sample statistics can help you further specify a parametric model.

Modelling a bivariate data



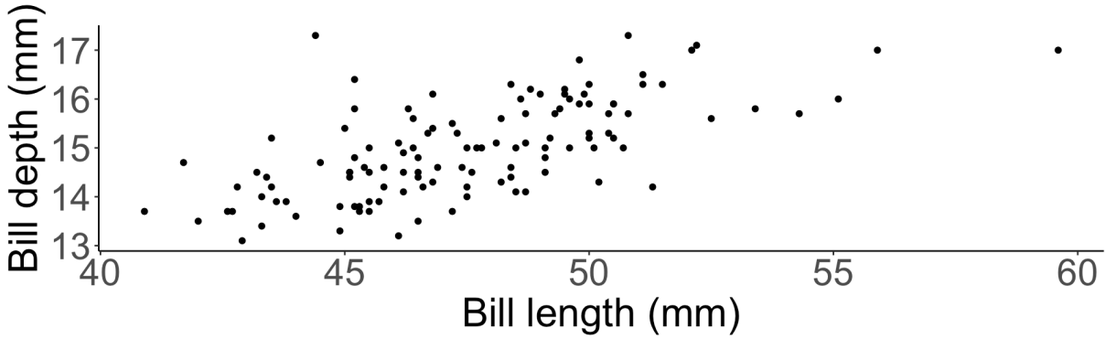
Note: In the raw data, bill dimensions are recorded as "culmen length" and "culmen depth". The culmen is the dorsal ridge atop the bill.

Artwork by @allison_horst.

Example: Gentoo penguins

Recall the data set on penguins collected at Palmer Station.

How would you model the *relationship* between bill depth and bill length?



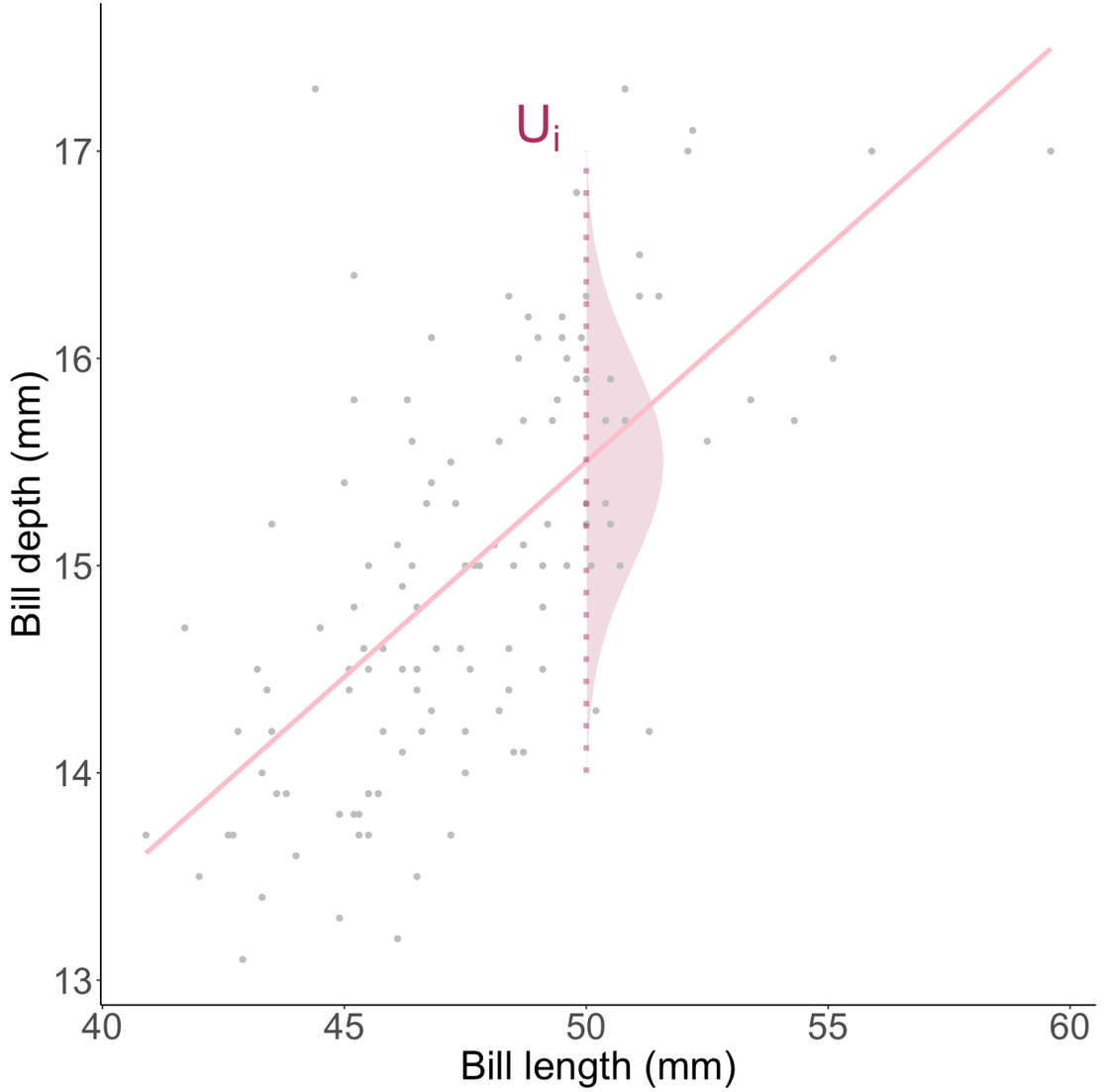
Modelling a linear relationship

A *deterministic* model, $y = g(x)$, can describe a linear relationship with

$$g(x) = \_\_\_\_\_$$

A statistical model for a linear relationship captures the randomness around the relationship

$$Y_i = g(x) + U_i$$



Simple linear regression model

- x-variable: **explanatory variable** or **independent variable**
- y-variable: **response variable** or **dependent variable**
- $y = \alpha + \beta \cdot x$: **regression line**
  - α: _____
  - β: _____

Consider bivariate data (x_1, y_1), (x_2, y_2), ..., (x_n, y_n). A **simple linear regression model** assumes that x_1, x_2, ..., x_n are nonrandom and that y_1, y_2, ..., y_n, are realizations of random variables Y_1, Y_2, ..., Y_n such that

$$Y_i = \alpha + \beta \cdot x_i + U_i$$

for $i = 1, 2, \ldots, n$ where U_1, U_2, ..., U_n are independent random variables with $\mathbb{E}(U_i) = 0$ and $\mathrm{Var}(U_i) = \sigma^2$.

Simple linear regression model

- x-variable: **explanatory variable** or **independent variable**
- y-variable: **response variable** or **dependent variable**
- $y = \alpha + \beta \cdot x$: **regression line**
  - α: _____
  - β: _____

> Y_i are independent, but not identically distributed.

Consider bivariate data (x_1, y_1), (x_2, y_2), ..., (x_n, y_n). A **simple linear regression model** assumes that x_1, x_2, ..., x_n are nonrandom and that y_1, y_2, ..., y_n, are realizations of random variables Y_1, Y_2, ..., Y_n such that

$$Y_i = \alpha + \beta \cdot x_i + U_i$$

for $i = 1, 2, \ldots, n$ where U_1, U_2, ..., U_n are independent random variables with $\mathbb{E}(U_i) = 0$ and $\mathrm{Var}(U_i) = \sigma^2$.

Example: Orange trees

- `Orange` in R contains a data set with 35 measurements of orange tree circumferences (mm) and the age of the trees in days at the time of measurement.
- Assuming these measurements are independent, a reasonable linear regression model would be

$$W_i = \alpha + \beta \cdot t_i + U_i$$

for $i = 1, \ldots, 35$ where

W_i is _____ and

t_i is _____ for ith measurement.

Example: Orange trees

- `Orange` in R contains a data set with 35 measurements of orange tree circumferences (mm) and the age of the trees at the time of measurement.
- Assuming these measurements are independent, a reasonable linear regression model would be

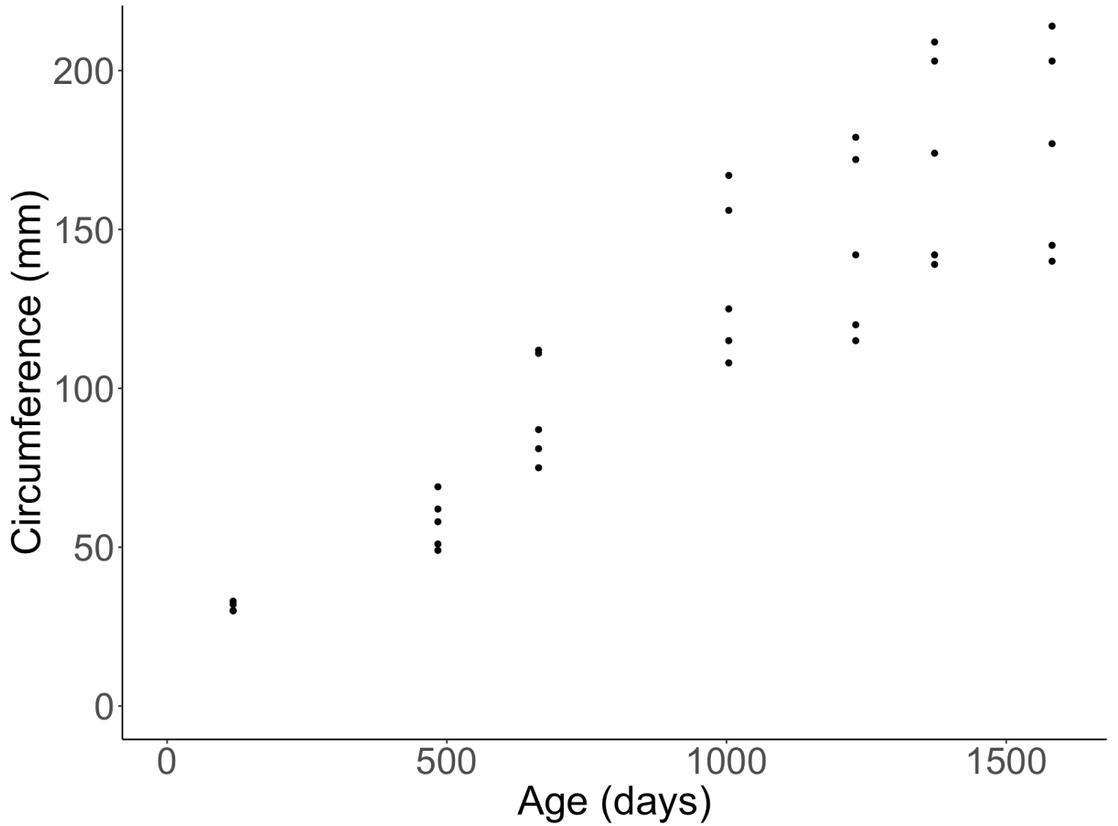
$$W_i = \alpha + \beta \cdot t_i + U_i$$

for $i = 1, \ldots, 35$ where

W_i is _____ and

t_i is _____ for ith measurement.

```r
# Orange data set is available from R.
Orange |>
  ggplot(aes(x = age, y = circumference)) +
  theme_classic() +
  geom_point() +
  # adjust axes limits
  coord_cartesian(xlim = c(0, 1600), ylim = c(0
  labs(x = "Age (days)", y = "Cicumference (mm)
```



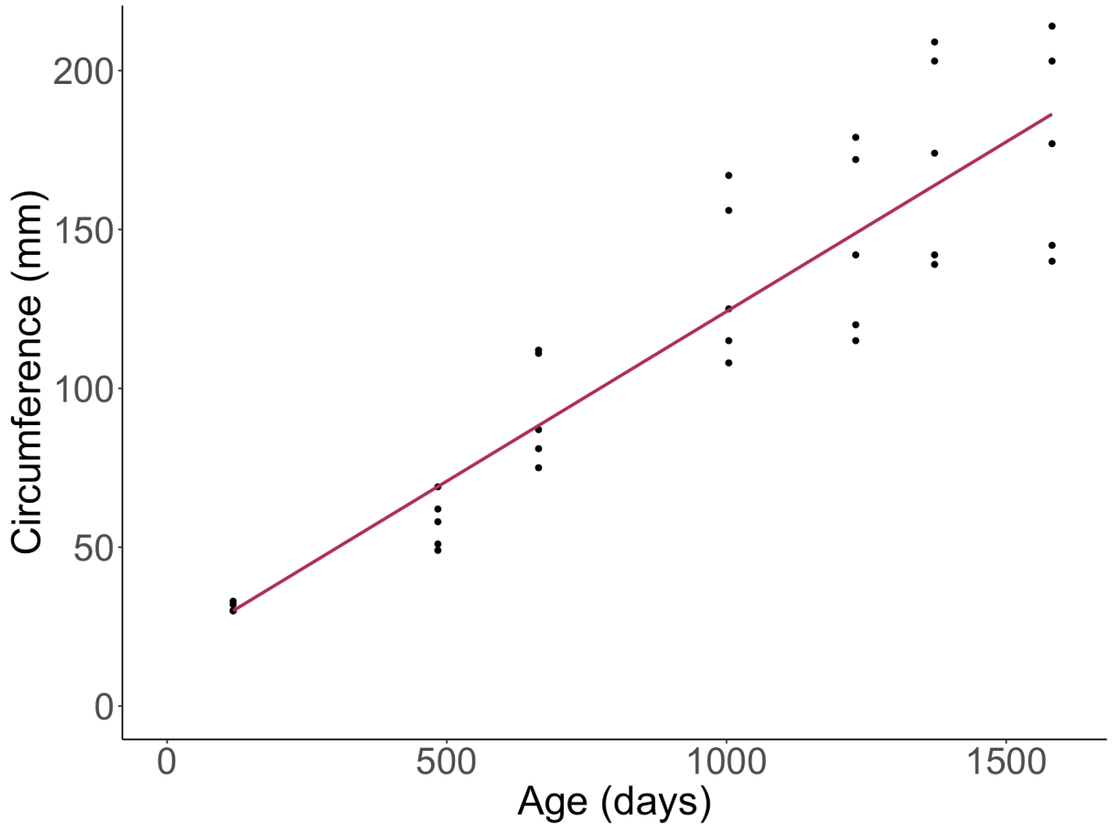
Example: Orange trees

Suppose the regression line that *best* describe the data is

$$y = 17.4 + 0.107 \cdot x.$$

What is the expected circumference of an orange tree that is 300 days old based on the model?

```
# Orange data set is available from R.
Orange |>
  ggplot(aes(x = age, y = circumference)) +
  theme_classic() +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, colour
  coord_cartesian(xlim = c(0, 1600), ylim = c(0
  labs(x = "Age (days)", y = "Circumference (mm
```



Example: Orange trees

Suppose the regression line that *best* describe the data is

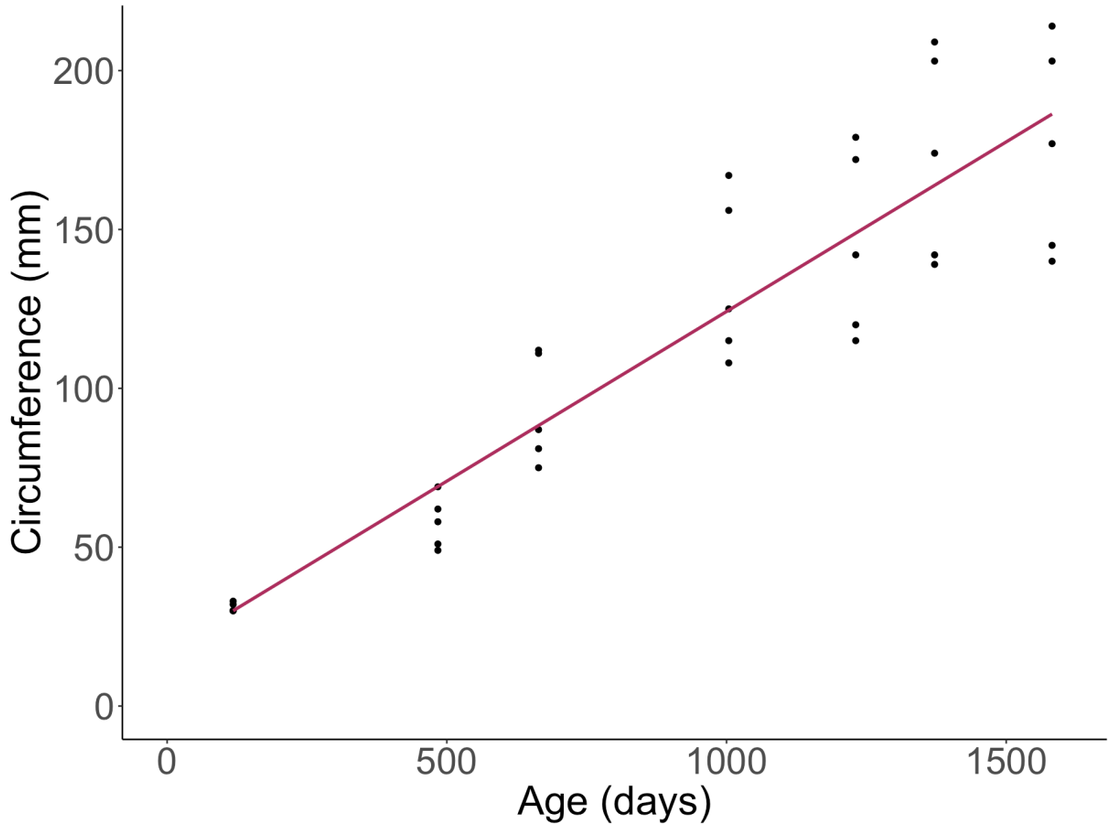
$$y = 17.4 + 0.107 \cdot x.$$

What is the expected circumference of an orange tree that is 300 days old based on the model?

```
1  17.4 + 0.107 * 300
```

```
[1] 49.5
```

```
1  # Orange data set is available from R.
2  Orange |>
3    ggplot(aes(x = age, y = circumference)) +
4    theme_classic() +
5    geom_point() +
6    geom_smooth(method = "lm", se = FALSE, colour
7    coord_cartesian(xlim = c(0, 1600), ylim = c(0
8    labs(x = "Age (days)", y = "Circumference (mm
```



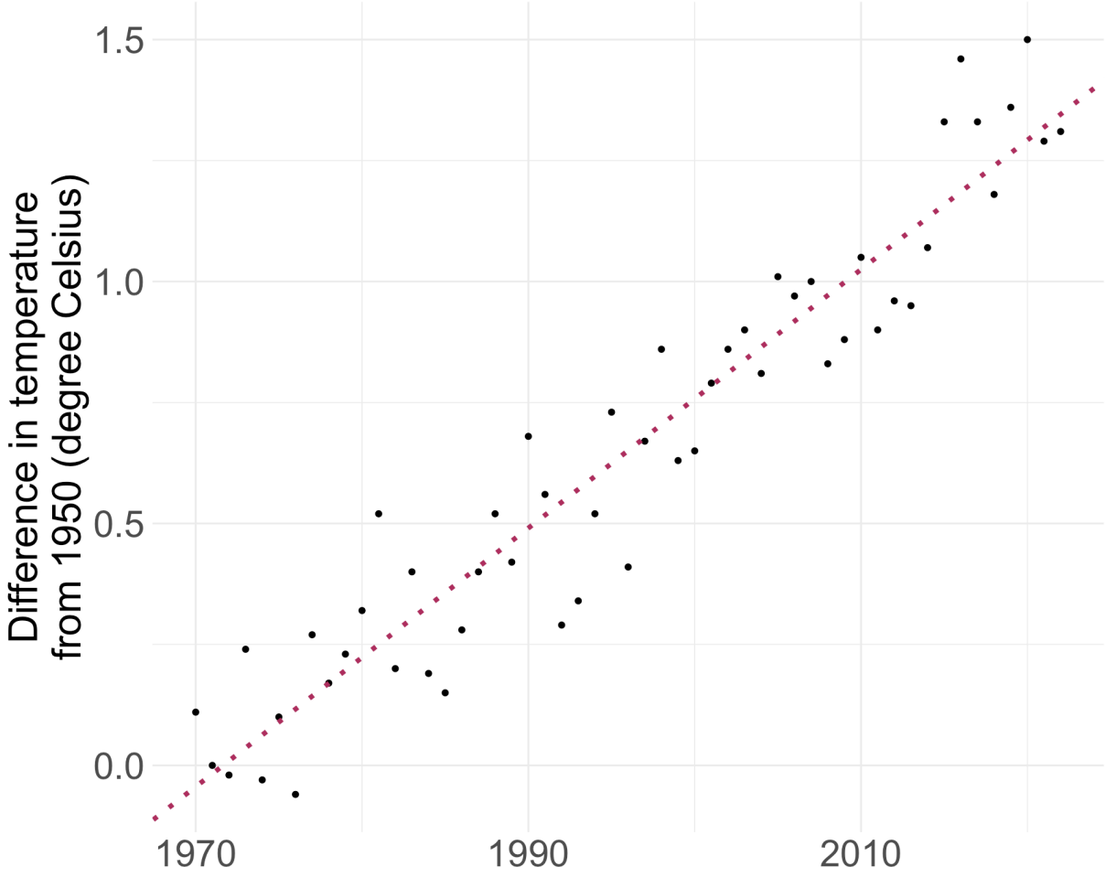
Example: Global surface temperature

Data on the right are annual averages of global surface temperature from 1970 to 2022 represented as the change from 1950. The dotted line is a regression line for

$$t_i = \alpha + \beta \cdot y_i$$

for $i = 1, \ldots, 53$ where y_i's are years from 1970 to 2022 and t_is are the recorded average global surface temperatures in year y_i.

Does year number explain the global temperature?



Data retrieved via tidytuesday with original data from NASA Goddard Institute for Space Studies at https://data.giss.nasa.gov/gistemp/.

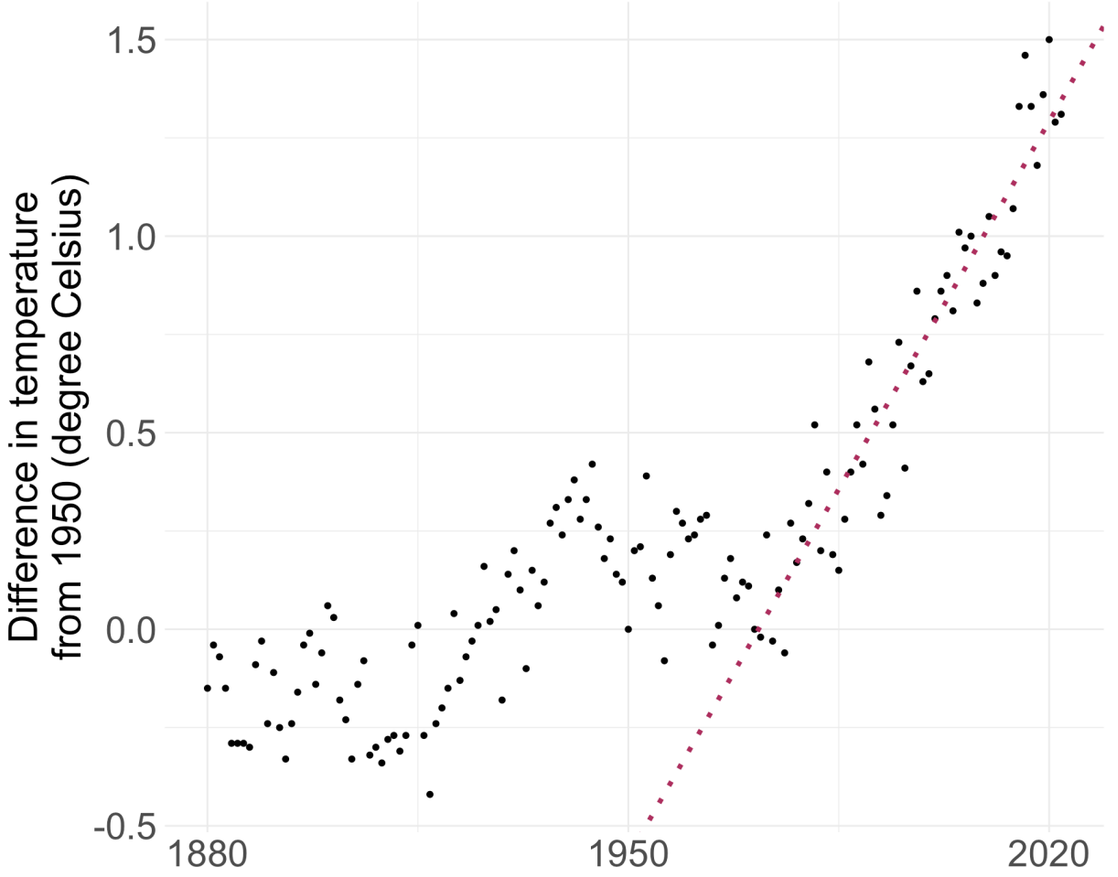
Example: Global surface temperature

Definitely not in the long run.

While the regression line may help describe the rate of change during the past 5 decades, it isn't suitable for describing the relationship beyond the time period.

Year doesn't explain global surface temperature. It is much more likely other factors contribute to the change in temperature.

Misspecified models lead to misleading results.



Data retrieved via tidytuesday with original data from NASA Goddard Institute for Space Studies at https://data.giss.nasa.gov/gistemp/.

Summary

- Parametric models can often be fully approximated using a few sample statistics.
- Simple linear regression models describe linear relationship between a nonrandom quantity and a random quantity as a sum of a regression line and a random effect variable.
- Unjustified models can lead to misleading explanations.