Welcome to Statistical Modelling I

CHRIS SUTTON, JANUARY 2024

Chris Sutton FIA FHEA

Email	c.sutton@qmul.ac.uk
Office	MB-B22

Office hours: Monday 12 – 1, Thursday 2 – 3

or online by appointment

Module summary and objectives

Aims: This course introduces linear regression models

- Which can be used for modelling relations between different variables
- Simple and multiple linear regression models will be studied and employed
- The methods studied in the course will be implemented in R

Learning Outcomes

- express regression models as linear equations or in matrix form
- estimate parameters of simple linear regression models by least squares
- calculate confidence intervals and predictive intervals for predictions
- explain methods for selecting variables in multiple regression models
- explain the effects of outliers and collinearity and how to detect them
- interpret computer output of the above methods

Teaching

The teaching for this module will run a bit differently from the 3 lectures + 1 tutorial you might be used to ☺

3 timetabled hours per week (not 4)

1 hour lecture on a Monday at 2 [Chris]
1 hour lecture on Thursday at 11 [Lubna]
1 hour IT Lab (from week 2) generally on Monday or Thursday

Short videos posted each week to cover all the statistical theory

- each video will be < 15 minutes long
- which videos you need to watch and when will be clearly signposted in QM Plus and in the Monday lectures

3 components will fit together for all you need

campus lectures

short videos

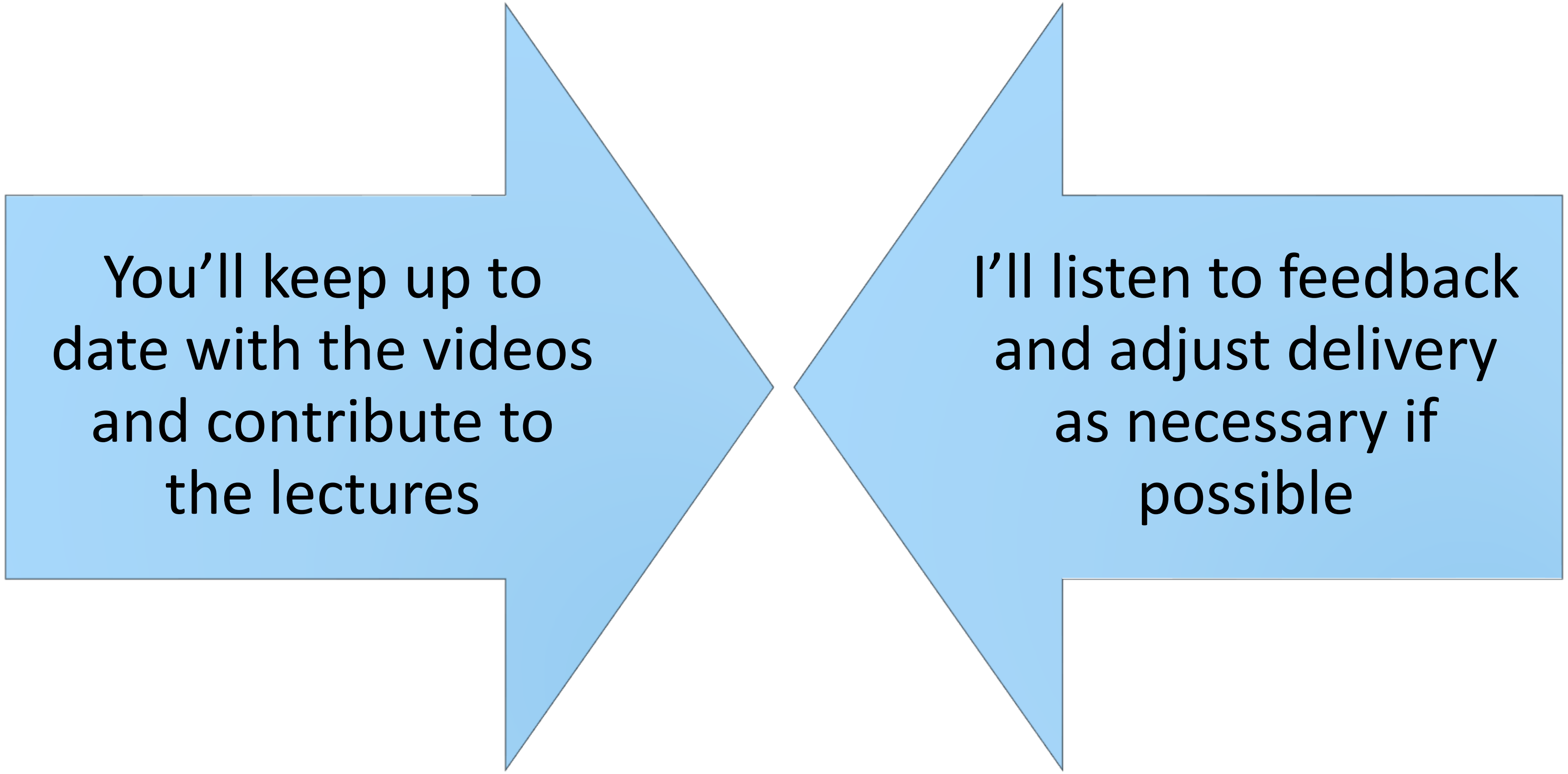
IT lab practice

Teaching content

This will be a mixture of:

- lectures with new statistics material
- discussions about real-world applications
- putting your own ideas into practice
- practice at problem questions
- working on example problems using R programming

We need to make an agreement here



You'll keep up to date with the videos and contribute to the lectures

I'll listen to feedback and adjust delivery as necessary if possible

Assessment

Assessment for this module

70%   Exam in May 2024

- handwritten, on campus
- 3-hour paper, answer every question

30%   assessed coursework using R set in 2 parts

- 1st part set in week 4 due in week 5
- there will be an initial step for this to complete by week 2
- 2nd part set in week 8 due in week 9

Institute and Faculty of Actuaries examinations (for Actuarial students only)

This module is one of three modules that combine to cover the material for the IFoA's CS1 "Actuarial Statistics" exam:

- Probability & Statistics II
- Statistical Modelling I
- Actuarial Statistics

For full details of how IFoA exemptions work see the page on QM Plus:

https://qmplus.qmul.ac.uk/mod/page/view.php?id=597978

Resources

The most important resources (and everything you will need to complete the module):

1. Your own lecture notes
2. Course materials posted on QM Plus
3. Practice questions in IT labs and exercise sheets
4. Additional online resources signposted
5. Independent study

Books

If you like to look at other resources to see how different people explain things try:

- **Gelman, Hill & Vehtari, *Regression and Other Stories* (CUP)**
- **Sheather, *A Modern Approach to Regression with R* (Springer)**

Other textbooks in the library:

- Sen & Srivastava, *Regression Analysis* (Springer)

- Draper & Smith, *Applied Regression Analysis* (Wiley)

- Weisberg, *Applied Linear Regression* (Wiley)

Topics in this Statistical Modelling module

1. Principles of statistical modelling
2. The Simple Linear Regression Model
3. Least Squares estimation
4. Properties of estimators
5. Assessing the model
6. Inference about the model parameters
7. Matrix approaches to simple linear regression
8. Multiple Linear Regression Models

3 objectives for our lectures and labs

- Introduce the mathematics of statistical modelling
- Construct and analyse some models using R
- Find our own real-world applications to talk about

Principles of Statistical Modelling

CHRIS SUTTON, JANUARY 2023

Principles of Statistical Modelling

Introductory topic describing modelling as an activity

The Simple Linear Regression Model

CHRIS SUTTON, JANUARY 2023

Topics in this Statistical Modelling module

1. Principles of statistical modelling
2. The Simple Linear Regression Model
3. Least Squares estimation
4. Properties of estimators
5. Assessing the model
6. Inference about the model parameters
7. Matrix approaches to simple linear regression
8. Multiple Linear Regression Models

The Model

Begin with a simple situation with

- one response variable, *Y*
- one explanatory variable, *X*

often,

- *X* can be controlled and is known
- *Y* is unknown but can be observed
- we have *n* pairs of observations { (x_1,y_1), (x_2,y_2), …, (x_n,y_n) }

Our modelling aim given n observations

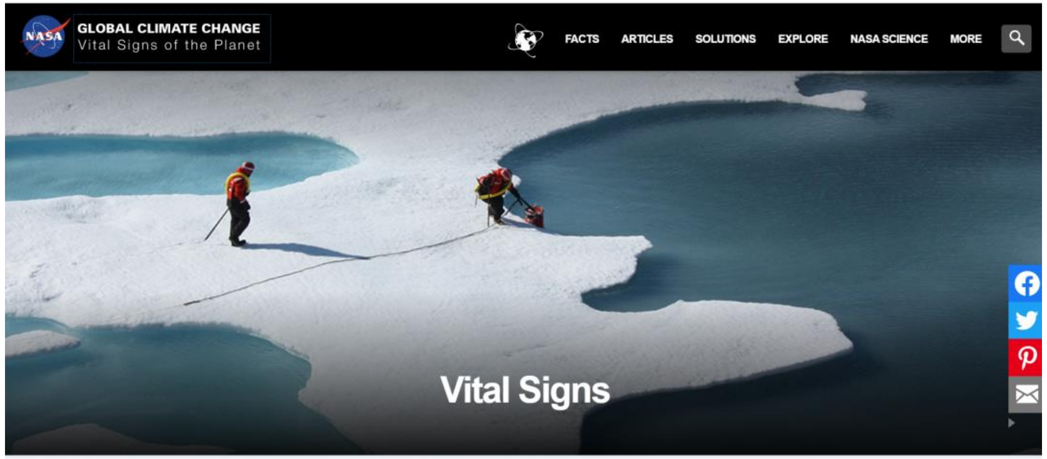
We would like to use these observations to estimate (or predict) the mean value of *Y* for some given values of *X*

A good place to start exploring the relationship between *X* and *Y* is a plot using the *n* pairs of observations

Global average temperature over time

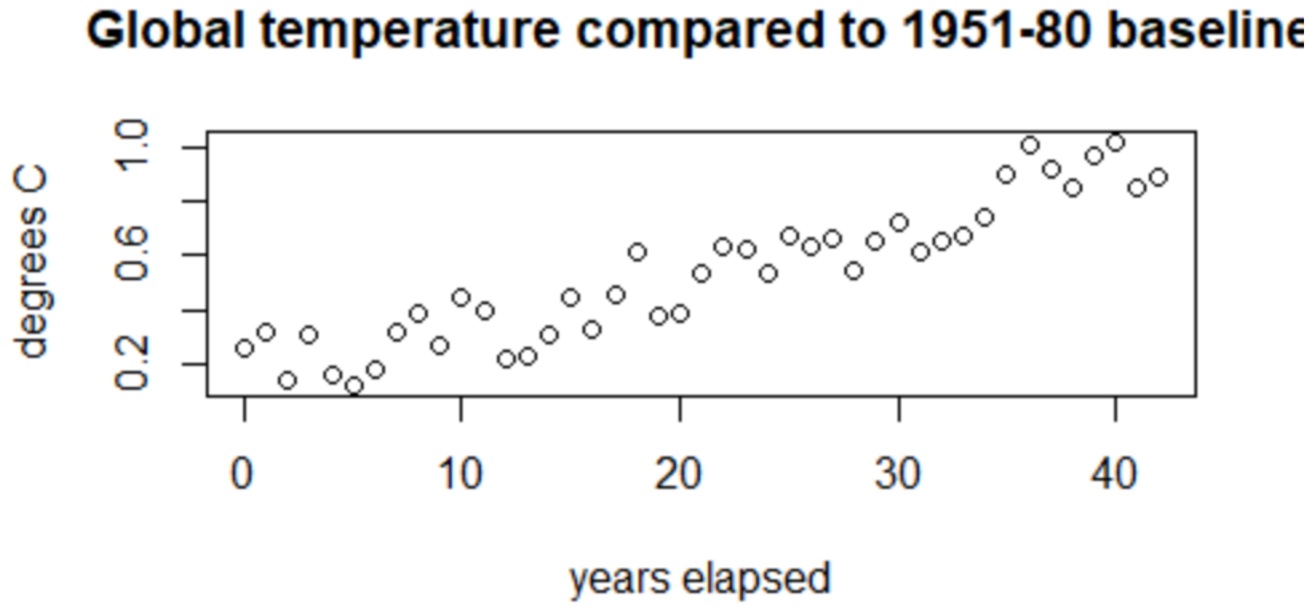
What is the evidence for climate change? How fast are average temperatures rising?

NASA Goddard Institute for Space Studies records average surface temperature each year and compares to a baseline temperature of the average for period 1951 - 1980



https://climate.nasa.gov/vital-signs/global-temperature/

Scatterplot of 1980 – 2022 data

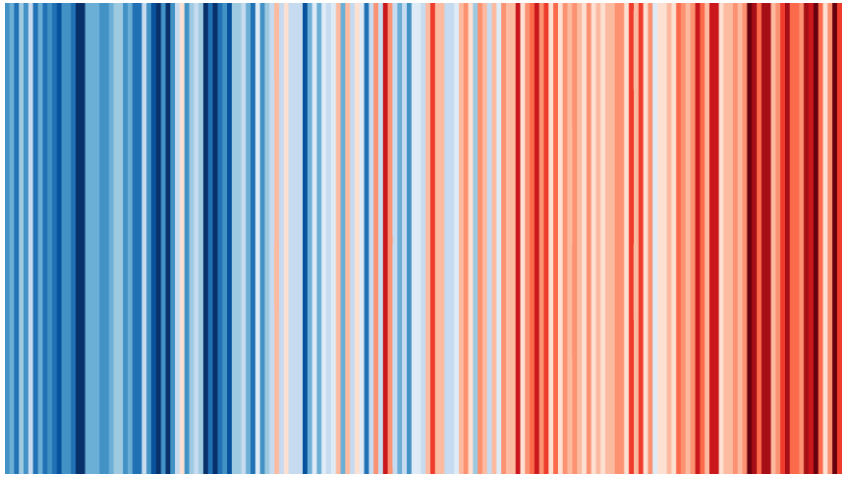


Your own statistical model

Think of 3 topics where you would be interested in building a (simple) statistical model:

1. An area you are interested in working in the future
2. Something you care about
3. An interest or hobby or something you support

My three



Parsimony

- Good practice to seek the simplest model that describes the relationship well
  - called the *principle of parsimony*
- What does “describes well” mean?
  - We will return to this issue a number of times through the module
- A *linear* relationship is the obvious place to start when looking for a simple model
  - This would be indicated by a plot being close to a straight-line

Linear (straight line) model

Given observation data (x_i, y_i) for i = 1, 2, … n we can fit a straight line to describe the response variable Y in terms of the explanatory variable X where

$$Y = \beta_0 + \beta_1 X$$

where,

- β_0 denotes the intercept
- β_1 is the slope of the line

Deterministic versus Stochastic

- This model is *deterministic*
  - it does not allow for any randomness
  - unlikely this model properly describes data which includes some random elements

We introduce randomness by having a *probabilistic* or *stochastic* element where the model for *Y* has 2 parts:

- the value for Y we expect to observe for a given value of *X*
- an additional uncontrolled random value

Stochastic Linear Model

The stochastic linear model can be written either as

$$Y_i = E[Y_i | X = x_i] + \varepsilon_i$$

or as

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

for *i* = 1, 2, … *n*

Here ε_i is the *random error.*

The random error

We usually make 3 assumptions about the random error:

(1) $E[\varepsilon_i] = 0$ for all *i*

(2) $var[\varepsilon_i] = \sigma^2$ for all *i*

(3) $cov[\varepsilon_i, \varepsilon_j] = 0$ for all $i \neq j$

Assumptions about Y_i

Because ε_i is a random variable, $Y_i$must also be a random variable

So we can re-write the 3 assumptions in terms of $Y_i|X = x_i$ rather than ε_i

(1) $E[\, Y_i|X = x_i] = \mu_i = \ \beta_0 + \ \beta_1\, x_i$ for all *i*

(2) $var[\, Y_i|X = x_i] = \ \sigma^2$ for all *i*

(3) $cov\left[Y_i|X = x_i,\ Y_j|X = x_j\right] = 0$ for all *i* ≠ *j*

The assumptions in words

- the dependence of *Y* on *X* is linear
- the variance of *Y* at each value of *X* is constant and does not depend on x_i
- Y_i and Y_j are uncorrelated

Simple Linear Model

Instead of $Y_i|X = x_i$ we often use $y_i = (Y_i|X = x_i)$

then the simple linear model can be written as

$$E[y_i] = \beta_0 + \beta_1 x_i$$

and

$$var[y_i] = \sigma^2$$

Normal assumption

- Often convenient also to assume the conditional distribution of Y_i is Normal
- This is the *Normal Simple Linear Regression Model*
- We will be working with this model for the first 4 weeks of the module

3 ways to write this model

(A) $y_i \sim N(\mu_i, \sigma^2)$ where $\mu_i = \beta_0 + \beta_1 x_i$

(B) $y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$

(C) $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ where the ε_i are iid $\varepsilon_i \sim N(0, \sigma^2)$

Your modelling question

Take one of the three areas you thought of earlier

Start to think of a question that you would like to answer with a (simple) statistical model

Question type is “what is the relationship between *x* and *y* ?”

Make it something that genuinely interests you

If you have time

Narrow the question down into a variable you would like to know more about and another variable that might be explanatory

My three

Do shares of high ESG rated smaller companies outperform in particular?

How quickly does habitat loss or deforestation lead to bird species extinction?

What are the best factors for predicting how good a season it will be for the BlueJays?