

# SMSTC: Probability and Statistics

Fraser Daly

Heriot-Watt University

September 2021

- Probability and Statistics
- Course outlines and teaching teams
- Prerequisites
- Assessment
- Feedback

“.. the true logic for this world is the calculus of Probabilities, which takes account of the magnitude of the probability which is, or ought to be, in a reasonable man's mind.”

– James Clerk Maxwell (1850)

*From the book “Probability theory: the logic of science” by E.T.Jaynes*

“.. the true logic for this world is the calculus of Probabilities, which takes account of the magnitude of the probability which is, or ought to be, in a reasonable man's mind.”

– James Clerk Maxwell (1850)

*From the book “Probability theory: the logic of science” by E.T.Jaynes*

Statistics may be defined as “a body of methods for making wise decisions in the face of uncertainty.”

– W.A. Wallis

- mathematical modelling of uncertainty: random events and random processes evolving in time
- crucial to understand dependence between different elements of the model
- in practice, driven by understanding properties of experimental observations
- correct measure of uncertainty of the decision making.

## Aims

- Building and analysing mathematical models of randomness, using elements of measure theory, functional analysis, combinatorics.
- Models include parameters, which can be specified in particular applications.

## Courses

- Foundations of Probability (Semester 1)
- Stochastic Processes (Semester 2)

# Foundations of Probability (Semester 1)

A gambler starts with  $\pounds X_0$ . At turn  $n = 1, 2, \dots$ , he stakes  $\pounds S_n$ , and

- gains  $\pounds S_n$  with probability  $p > 1/2$ , or
- loses  $\pounds S_n$  with probability  $1 - p$ .

We let  $\pounds X_n$  be his total wealth after turn  $n$ , and assume (reasonably!) that  $0 \leq S_n \leq X_{n-1}$ .

How can the gambler maximize his long-term gain?

Calculations using *conditional expectation* show that  $E(X_n)$ , the gambler's average wealth after turn  $n$ , is maximised by choosing  $S_n = X_{n-1}$ . But, this is not a viable long-term strategy (what happens the first time you lose?)...

# Foundations of Probability (Semester 1)

If we instead try to maximise  $E \log(X_n)$ , we can show that this is achieved using the strategy  $S_n = (2p - 1)X_{n-1}$ .

One way to do this is to show that a certain linear shift of  $\log(X_n)$  is a *martingale* in this case, and a *supermartingale* in all others.

We can also check, using the *law of large numbers*, that if

- our gambler uses this strategy, and has  $\pounds X_n$  after turn  $n$ , and
- another gambler uses the strategy  $\tilde{S}_n = \lambda \tilde{X}_{n-1}$  (where  $\lambda < 1$  and  $\lambda \neq 2p - 1$ ), and has  $\pounds \tilde{X}_n$  after turn  $n$

then  $X_n/\tilde{X}_n$  grows exponentially for large  $n$ , with probability 1. Hence, the choice  $\lambda = 2p - 1$  is a better choice than any other.



# Foundations of Probability (Semester 1)

- **Fundamentals:** probability spaces,  $\sigma$ -algebras, probability measures, conditioning and independence
- **Random variables** and their distributions, important special distributions (binomial, Poisson, geometric, normal, exponential etc.)
- **Convergence** and **limit theorems**
- **Conditional expectation** and **martingales**
- **Renewal theory**

# Stochastic Processes (Semester 2)

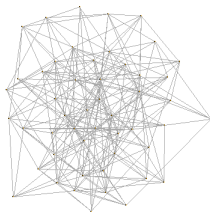
Suppose we have  $n$  vertices/nodes.

Each pair of vertices is joined by an edge/link with probability  $p$ , independently of all other pairs of vertices.

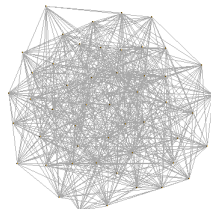
This is the Erdős–Rényi random graph  $G(n, p)$ . It can be used to model a ‘typical’ (or ‘unstructured’ or ‘random’) communication (or power, or distribution, or ...) network, for example.



$p = 0.05$



$p = 0.2$



$p = 0.5$

Let  $p = c/n$ . Then (under some mild conditions on  $c$ )  $G(n, p)$  contains a path of length at least constant  $\times n$  with probability 1, for large enough  $n$ .

This is proved by analysing an algorithm which explicitly constructs such a path, and exploiting the *Markovian* structure present in the algorithm.

# Stochastic Processes (Semester 2)

Let  $K_n$  be the complete graph, with  $n$  vertices and an edge between each pair of vertices. Suppose we colour each edge of  $K_n$  either red or blue.

There is a colouring of  $K_n$  which contains at most  $\binom{n}{a} 2^{1-\binom{a}{2}}$  monochromatic copies of the complete graph  $K_a$ .

# Stochastic Processes (Semester 2)

Let  $K_n$  be the complete graph, with  $n$  vertices and an edge between each pair of vertices. Suppose we colour each edge of  $K_n$  either red or blue.

There is a colouring of  $K_n$  which contains at most  $\binom{n}{a}2^{1-\binom{a}{2}}$  monochromatic copies of the complete graph  $K_a$ .

We can prove this by

- Randomly colouring  $K_n$  (each edge is red with probability  $1/2$ , or blue otherwise, independently of the other edges);
- Calculating that the average number of monochromatic copies of  $K_a$  is  $\binom{n}{a}2^{1-\binom{a}{2}}$ ; and
- Concluding that there must exist a colouring with at most this many monochromatic copies of  $K_a$ .

- **Markov chains** and **processes**, **Poisson processes**
- **Applications**, including connections to statistics and graph theory
- **Brownian motion** and **stochastic calculus**

- Yvain Bruned (Edinburgh)
- Burak Buke (Edinburgh)
- Fraser Daly (Heriot-Watt)
- Sergey Foss (Heriot-Watt)
- István Gyöngy (Edinburgh)
- Mateusz Majka (Heriot-Watt)

For questions related to the Probability modules, feel free to get in touch with Fraser Daly ([f.daly@hw.ac.uk](mailto:f.daly@hw.ac.uk)).

Tutorial support will be arranged locally.

- Elements of mathematical analysis, linear algebra and combinatorics at undergraduate level.
- For Stochastic Processes, in addition: Probability theory, either at undergraduate level or from Foundations of Probability.



Each module is assessed by two written assignments.

Approximate deadlines:

- Foundations of Probability: mid-November and early January.
- Stochastic Processes: mid-February and end of March.

Assignments will be available at least two weeks before the deadline.

Solutions for (at least) one assignment from each module should be prepared using  $\text{\LaTeX}$ .

## Aims

- Model fitting from experimental data: How do we select an appropriate model? How do we fit parameters to a given data set? How do we handle imperfect (missing/contaminated/...) data? How do we quantify uncertainty in our estimates?
- Testing plausibility of given conjectures.
- Simulation of intractable probability distributions.

## Courses

- Regression and Simulation Methods (Semester 1)
- Modern Regression and Bayesian Methods (Semester 2)

# Regression and Simulation Methods (Semester 1)

Linear model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i,$$

for  $i = 1, \dots, n$  (where  $n$  is the sample size), and where  $\epsilon_1, \dots, \epsilon_n$  are independent and identically distributed with  $\epsilon_1 \sim N(0, \sigma^2)$ .

More succinctly

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}).$$

Residual Sum of Squares:

$$\text{RSS} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}),$$

minimized by choosing

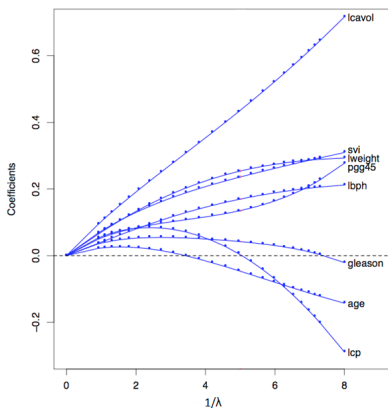
$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}.$$

# Regression and Simulation Methods (Semester 1)

What happens when  $\mathbf{X}^T \mathbf{X}$  is singular?

One possible solution: Ridge regression

$$\hat{\beta}^{\text{ridge}} = (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^T \mathbf{y}.$$



- **Introduction to R**
- **Linear models:** Estimation, testing, model checking, factors, model fitting in R. Analysis of simple designed experiments. Case studies.
- **Likelihood and optimisation:** Likelihood principles and key distributional results. Examples. Newton's method for optimisation. Two-parameter likelihoods. General optimisation methods. Implementation in R.
- **Generalised linear models:** Exponential family. Link functions. Examples. Iteratively weighted least squares. Model fitting in R. Case studies.
- **Simulation and bootstrapping:** Non-parametric bootstrap; confidence intervals; implementation in R. Parametric bootstrap. Simulation methods and implementation in R.
- **Case study**

# Regression and Simulation Methods (Semester 1)

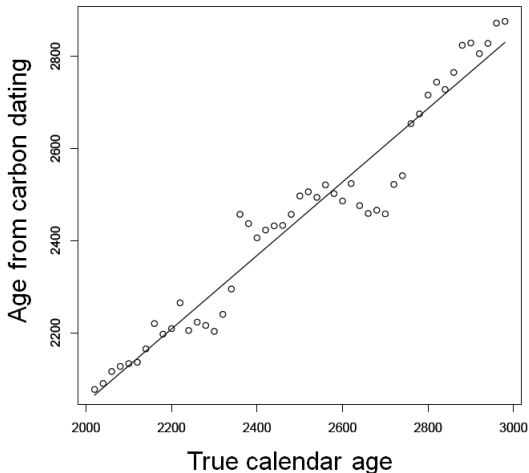
The first half of Regression and Simulation Methods will be run as an online audio/video course. It covers what for many will be revision, and this flexible form of delivery allows participants to study different parts of the material at a speed and depth appropriate for them.

We ask you to check the course materials on the SMSTC website. If any of it is unfamiliar, you can view the relevant lectures, and attempt the related tutorial questions.

Regular Zoom sessions will begin in the sixth session (9th November).

# Modern Regression and Bayesian Methods (Semester 2)

Radiocarbon data: high precision measurements of Carbon-14 in Irish oak, used to construct a calibration curve (here with line of best fit)

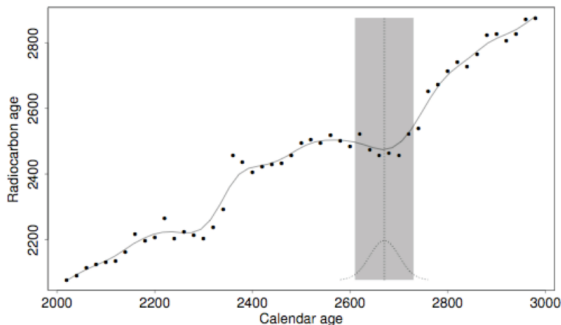


# Modern Regression and Bayesian Methods (Semester 2)

One solution to non-linearity: *local* linear regression. Solve

$$\min_{\alpha, \beta} \sum_{i=1}^n \{y_i - \alpha - \beta(x_i - x)\}^2 w(x_i - x; h),$$

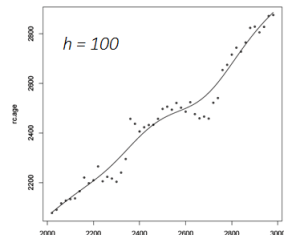
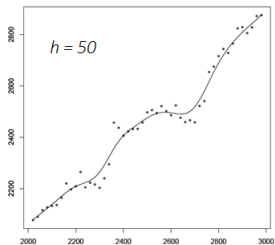
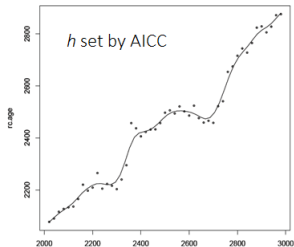
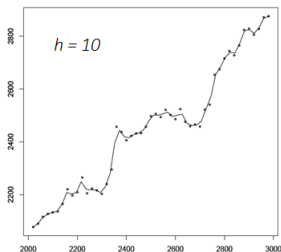
for a weight function  $w$ , and take  $\hat{\alpha}$  as the estimate at  $x$ .





# Modern Regression and Bayesian Methods (Semester 2)

We have a choice of the parameter  $h$ :



# Parameter estimation framework

Let  $\mathbf{Y} = (Y_1, \dots, Y_n) \sim p(\cdot | \theta)$  – likelihood, for some  $\theta \in \Theta \subseteq \mathbb{R}^p$ .

**Bayesian model and Bayes estimator of  $\theta$ :**

**Prior distribution:**  $\theta \sim p(\theta)$ ,  $\theta \in \Theta$  – density of the prior distribution.

**Posterior distribution:**  $p(\theta | \mathbf{Y}) \propto p(\mathbf{Y} | \theta) p(\theta)$ ,  $\theta \in \Theta$ .

**Bayes estimator :** for a given a loss function  $Q(\hat{\theta}, \theta)$ ,

$$\hat{\theta} = \arg \min_{\hat{\theta} \in \Theta} \int Q(\hat{\theta}, \theta) p(\theta | \mathbf{y}) d\theta.$$

E.g.

- $Q(x, y) = (x - y)^2$  - posterior mean  $\hat{\theta} = E(\theta | \mathbf{y})$
- $Q(x, y) = |x - y|$  - posterior median
- $Q(x, y) = I(x = y)$  - maximum a posteriori estimator (MAP)

For an appropriate loss function, it can be a decision, credible interval/region ..

# Statistical inference for high dimensional data

**Likelihood:**  $\mathbf{Y} = (Y_1, \dots, Y_n) \sim p(\mathbf{Y} \mid \theta)$ , for some  $\theta \in \Theta \subseteq \mathbb{R}^p$ ,  $p \gg n$ .

**Aim:** to estimate unknown  $\theta$ , its confidence region, make decisions.

# Statistical inference for high dimensional data

**Likelihood:**  $\mathbf{Y} = (Y_1, \dots, Y_n) \sim p(\mathbf{Y} | \theta)$ , for some  $\theta \in \Theta \subseteq \mathbb{R}^p$ ,  $p \gg n$ .

**Aim:** to estimate unknown  $\theta$ , its confidence region, make decisions.

## Penalised log likelihood

**estimator:**

$$\hat{\theta} = \arg \min_{\hat{\theta}} \left[ -\log p(\mathbf{Y} | \hat{\theta}) + \text{pen}(\hat{\theta}) \right]$$

where penalty reflects desirable properties of the solution, e.g. sparsity.

**Problems:**

- Construction of confidence regions for  $\hat{\theta}$  and other decision making.
- Assumptions of theoretical guarantees are often not verifiable.

## Bayesian model:

Given prior  $p(\theta)$ , **posterior distribution** is

$$p(\theta | \mathbf{Y}) = \frac{p(\mathbf{Y} | \theta) p(\theta)}{\int_{\Theta} p(\mathbf{Y} | \theta) p(\theta) d\theta},$$
$$\hat{\theta} = \arg \max_{\hat{\theta}} E \left( Q(\hat{\theta}, \theta) | \mathbf{Y} \right)$$

given a loss function  $Q$  on  $\Theta \times \Theta$ . Bayesian analogues of a confidence region and decision making can be constructed.

# Modern Regression and Bayesian Methods (Semester 2)

- **Random effects models:** Methods for linear and non-linear mixed effects models. Case studies.
- **Modern regression:** Density estimation. Non-parametric regression. Bandwidth selection. Examples. Additive models. The backfitting algorithm. Examples.
- **Bayesian methods:** Priors and posteriors. Prior sensitivity. Marginal distributions.
- **Markov chain Monte Carlo:** Metropolis-Hastings algorithm. Gibbs sampler. Convergence, burn-in, mixing properties, tuning parameters. WinBUGS. MCMC simulations in R. Examples. Advanced topics: eg, random effects, missing data, model selection.
- **Case study**

Semester 1 (Regression and Simulation Methods):

- Weeks 1-5: Adrian Bowman (Glasgow) - as podcasts
- Weeks 6-10: Victor Elvira (Edinburgh)

Semester 2 (Modern Regression and Bayesian Methods):

- Weeks 1-5: Adrian Bowman (recorded lectures) (Glasgow)  
+ TBD (live office hour) (Glasgow)
- Weeks 6-10: TBD (St Andrews)

For questions related to the Statistics modules, feel free to get in touch with Victor Elvira ([victor.elvira@ed.ac.uk](mailto:victor.elvira@ed.ac.uk)).

Tutorial support will be arranged locally.

- Basic concepts in probability (elementary probability distributions), statistics (idea of estimation, confidence intervals, hypothesis tests), calculus, and linear algebra. These would usually be provided in first undergraduate courses.
- For Modern Regression and Bayesian Methods: the semester 1 course (Regression and Simulation Methods), or equivalent.

## Regression and Simulation Methods:

- One written assignment (based on the final five lectures), deadline in early-mid January. The assignment will be available by mid-December.

## Modern Regression and Bayesian Methods:

- Two written assignments, one after each block of five lectures (around mid February and end of March). Assignments will be available at least two weeks before the deadline.



- if you have any questions/concerns, get in touch with me or another member of the teaching team.
- feedback and questions are encouraged during lectures.
- please don't wait for the end of the module!