UNIVERSITY
OF HULL

Statistical Models (551305)

Trimester 2

Silvio Fanzon

Email: S.Fanzon@hull.ac.uk

Motivation

- Statistical Models builds on Y1 module **Intro to Prob & Stats**
- Teaches you the statistical computer language R
  (widely used in research as well as in industry)
- Useful for project work and dissertations
- Statistical modelling and simulation widely applied
  (sports, finance, LLMs - such as ChatGPT, etc.)
- Good foundation for Data Science oriented careers

Module content

- Normal distribution family (t, F, χ^2)
- One-sample hypothesis tests
- Two-sample hypothesis tests
- Tests for contingency tables
- Regression Models

Example: Man Utd performance

Manager	Won	Drawn	Lost
Moyes	27	9	15
Van Gaal	54	25	24
Mourinho	84	32	28
Solskjaer	91	37	40
Rangnick	11	10	8
ten Hag	61	12	28

Table: Performance of Man Utd managers since 2014

- Man Utd performance declined in the post-Sir Alex Ferguson era
- Since 2014 Man Utd has had 6 different managers (excluding interims)

Example: Man Utd performance

Manager	Won	Drawn	Lost
Moyes	27	9	15
Van Gaal	54	25	24
Mourinho	84	32	28
Solskjaer	91	37	40
Rangnick	11	10	8
ten Hag	61	12	28

Table: Performance of Man Utd managers since 2014

- **Question:** Are managers to blame for Man Utd performance? *(Spoiler: No)*
- Analysis possible using **Tests for Contingency Tables**

Real-World Scenario: Buying a House

1. You secure a mortgage with Halifax to buy a house
2. You find the perfect house in Hull, and agree to buy for 120,000 GBP
3. Your solicitor contacts Halifax to prepare the mortgage
4. Halifax needs to ensure the house is worth 120k before approval
5. This is to protect them in case of default, as they would repossess the house
6. After 5 minutes, Halifax deems the house worth 120k and approves the mortgage

Question: How is this possible without seeing the property?

Answer: Halifax uses a statistical model to predict house prices based on features like size, number of bedrooms, and postcode

We can build our own pricing model using Linear Regression

Linear Regression

Model used to analyze the (linear) relationship between

- a dependent variable Y (prediction)
- and one or more independent variables X_i (predictors)

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_n X_n + \varepsilon$$

Goal: Given values $X_1, \ldots, X_n \quad \rightsquigarrow \quad$ predict Y (up to error ε)

Example: Predicting Housing Prices

- $Y =$ House Price
- $X_1 =$ Size (in square feet)
- $X_2 =$ Number of bedrooms
- $X_3 =$ Postcode

Given Data: Prices of houses in Hull, with size, bedrooms, postcode

House Price (GBP)	Size (sq ft)	# Bedrooms	Postcode
150000	850	2	HU1
230000	1200	3	HU2
270000	1500	4	HU4

Phase 1. Fit the Model: Estimate parameters β_i for the model

$$Y = \beta_0 + \beta_1 \times \text{Size} + \beta_2 \times \text{Bedrooms} + \beta_3 \times \text{Postcode}$$

Parameters: Maximize the likelihood of observing the actual house prices

Example: With high probability, we must have

$$150000 \approx \beta_0 + \beta_1 \times 850 + \beta_2 \times 2 + \beta_3 \times 1$$

This has to hold for all the Houses in the table

Phase 2. Use the model: Input new features to predict price

These ideas are incredibly powerful

1. Fit a statistical model to data
2. Use the model for predictions

Example: ChatGPT learns a function f that predicts:

$$f(\text{Sentence}) = \text{Probability distribution over next words}$$

While f is more advanced than simple regression, the core idea remains:

- Learn parameters so f predicts the most likely next word(s) based on a dataset
- Use f for predictions (e.g., writing your assignments!)

Disclaimer: We are not going to cover such models (Neural Networks)

`https://writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work/`

Learning outcomes

- Statistical models for inference on given data sets
- Formulate and test hypotheses + interpret the results
- Linear Regression Models to analyse relations between variables
- Discuss assumptions underlying given statistical models
  Do such assumptions hold?

Module organization

Teaching: Each week we have

- 2 lectures of 2 hours
- 1 tutorial of 1 hour

Assessment:

- 10 problem sheets (accounts for 30% of final mark)
- Coursework (accounts for 70% of final mark)

Get in touch for more information:

Silvio Fanzon

S.Fanzon@hull.ac.uk