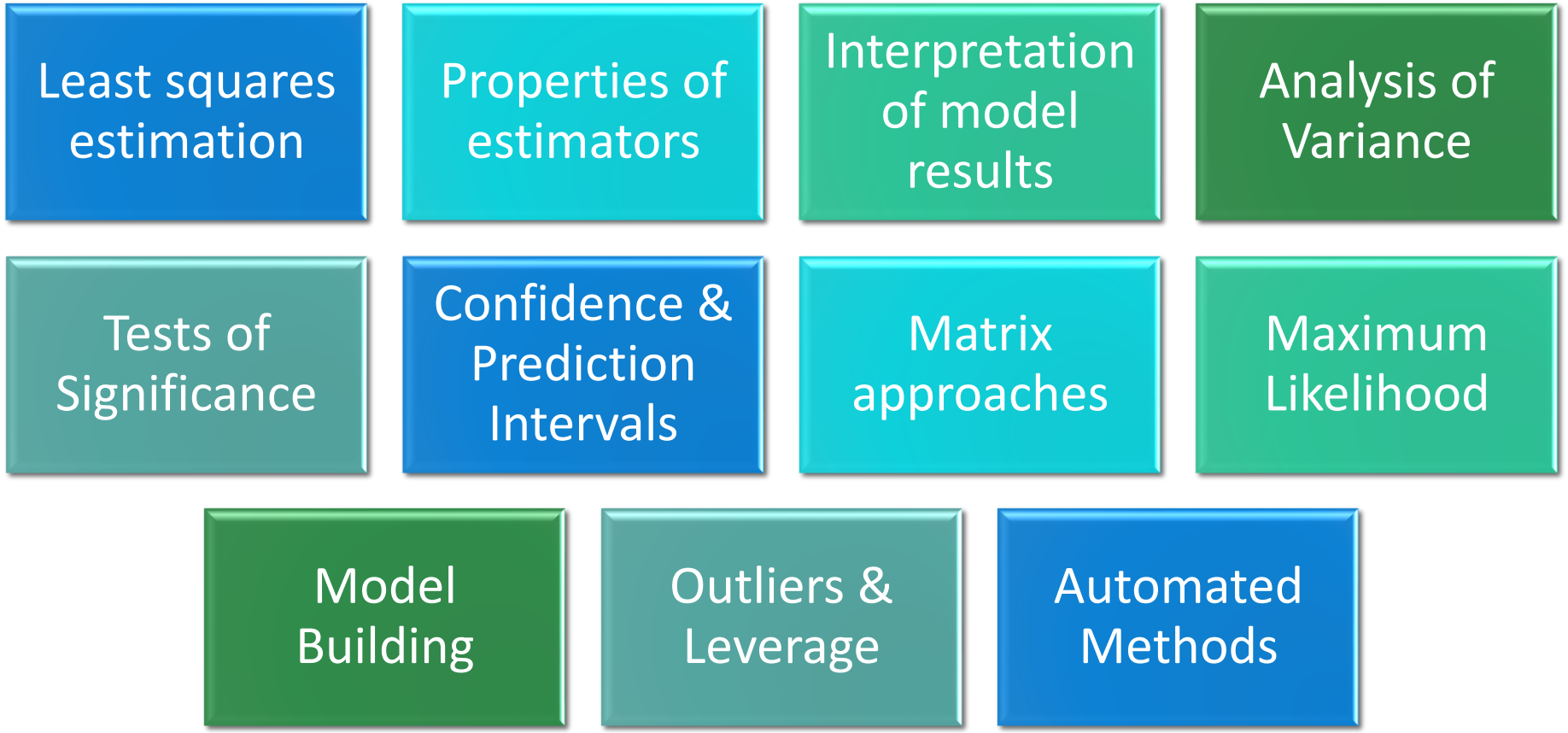
Revision lecture 1

CHRIS SUTTON, APRIL 2024

Overview of Statistical Modelling

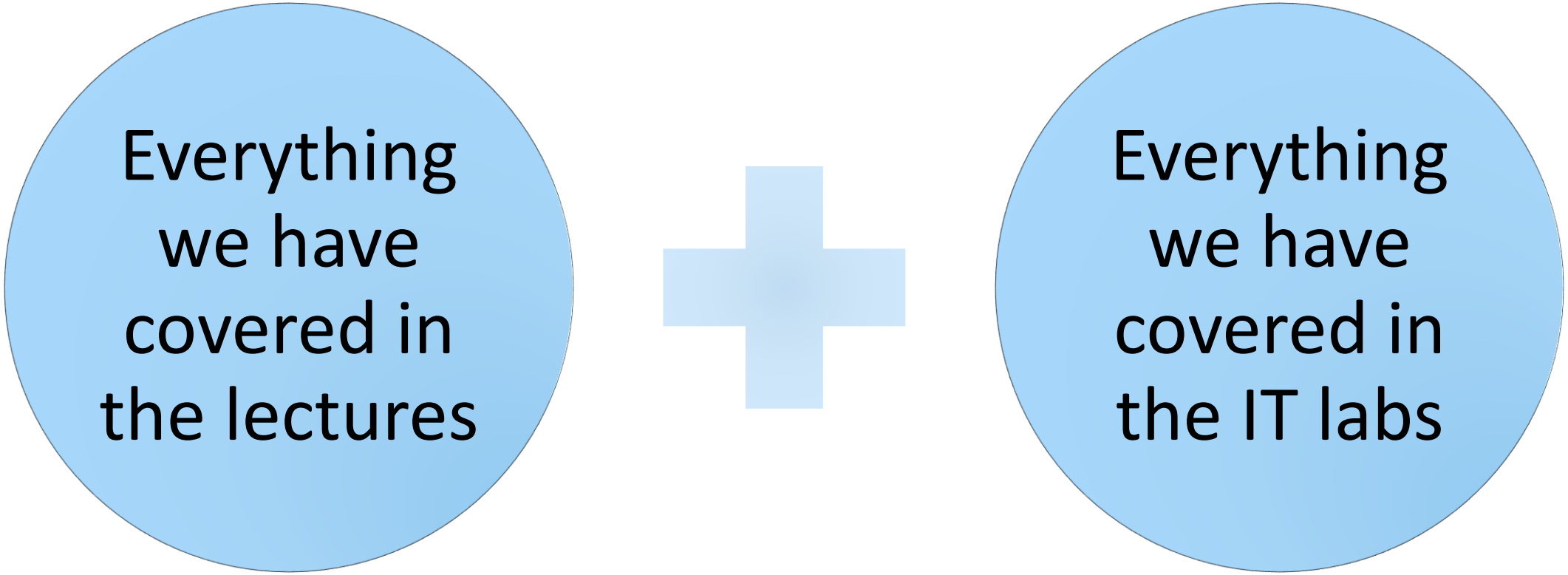


Exam information

May Exam

- Paper and pen on campus
- You will need a calculator, but you cannot bring notes
- 3 hours, 4 questions, answer every question
- 2 questions on simple linear regression, 2 questions on multiple linear regression
- Actuarial students - part of CS1 exemption

What material is examinable?



Types of Questions

Complete calculations to find LS estimates in SLRM	Calculations needed to construct ANOVA table	Tests of Hypothesis for significance of model or parameters
Calculations that use fitted model results	Calculations needed to evaluate suitability of data for modelling	Discussion of calculation results

Types of Questions continued

Commentary on variety of plots

Explain what some R code is doing

Explain steps needed in model analysis or checking

Re-state model in matrix form

Show how different matrices are used in linear models

Find a Maximum Likelihood Estimator

Types of Questions continued

Interpret model results and parameter values

Calculations for subset deletions or model building

Compare two methods by discussing their (dis)advantages

Explain the properties of different Statistics

Write R code needed to perform certain tasks

Explain desirable model properties

What you will need to be able to do

- interpret R output and use it to do your own calculations
- remember formulae from the lectures
- do tests of hypothesis properly
- explain how we analyse models and describe the steps taken
- explain how we select multiple regression models and describe the steps taken [remembering we covered different ways to do this]

What there won't be on the 2024 exam

- Proofs
- Trick questions

Multiple linear regression models

Multiple linear regression

Model construction and analysis is the same as the simple linear regression model

R code still `lm()` `anova()` `fitted()` `rstandard()` `plot()` etc

The key new thing compared to simple model is model selection

There are 2^{p-1} possible models – which ones are best?

Remember there is no one right answer, it depends on:

- Explanatory power versus Parsimony
- Which method is used for model selection

Model selection approaches

Method	Be able to do calculations	Be able to describe the process including remember formulae and R code needed
Deleting variables according to the extra sum of squares principle by F tests	√	√
Use of statistics (MS_E, R^2, Adjusted R^2, Mallows) to select from all subsets		√
Backwards elimination as an automated approach		√
Stepwise regression as an automated approach		√
Use of AIC in an automated approach		√

Matrix approaches

What you need to know about matrices

How to write linear regression models in matrix form

What each vector and matrix represents

The formula for the least squares estimate of vector $\boldsymbol{\beta}$ in terms of matrices $\mathbf{Y}$ and $\mathbf{X}$

The formula for the vector of fitted values

What the hat matrix is and how we use it

Why it is important that X^TX is invertible

Multicollinearity

Multicollinearity

What is it?

- where an X variable in a multiple regression model is a linear combination of one or more of the other X variables

Why are we concerned with this?

- it leads to estimates of betas with high variance (which means we won't have a lot of confidence in the reliability of our estimate or the model)

How do we check for it?

- calculate the Variance Inflation Factor and worry where VIF > 10