MATH3091 Statistical Modelling II
Lecture 1: Module Information

Dr. Chao Zheng and Prof. Sujit Sahu

Overview:

- The teaching of this module will be shared by Dr Chao Zheng and Prof Sujit Sahu
- Chao (during weeks 1-5) will teach the first part on Linear Models.
- This part will include review of Math2010 and introduction to new materials on likelihood based inference (e.g. maximum likelihood estimation) and linear mixed models.
- Sujit (during weeks 6-10) will introduce the generalised linear models for modelling count data (e.g. number of Covid-19 deaths in England).
- There will be approximately 36 contact hours scheduled (see the next slide) – all during weeks 1-10.

Plan of the 36 contact hours in MATH3091

- 22 lectures
  - 2 or 3 lectures per week on Monday and Friday
  - video recordings from last year will be available to watch
- 8 computer labs
  - Monday 17:00-18:00
- 6 problem classes
  - usually on Fridays
  - problem sheets will be provided much ahead of time.
  - you are asked to attempt the problems beforehand to get more out of these sessions.

Arrangement (Weeks 1-5)

- Chapter 1: Preliminaries
- Chapter 2: Likelihood Based Statistical Theory
- Chapter 3: Linear models
- Chapter 4: Linear Mixed Models

	Monday (am)	Monday (pm)	Friday (pm)
Week 1	Lec 1	Lab 1(preparation)	Lec 2, Lec 3
Week 2	Lec 4	Lab 2(Chap 2)	Lec 5, PC 1(Chap 2)
Week 3	Lec 6	Lab 3(Chap 3)	Lec 7, PC 2(Chap 3)
Week 4	Lec 8	Lab 4(Chap 4)	Lec 9, PC 3(Chap 4)
Week 5			***Class Test***

- We might have a Lec 9' and Lab 4' in the Monday of Week 5, depending on our progress

Assessment methods and key dates

- 1. **In-person Class Test: — 25%**
  - Time: **4-6 PM on March 4**
  - **It cannot be taken at any other date, time or in other format.**
  - It will be based on the contents taught during Weeks 1-5.
  - It is open book but you cannot use any electronic devices (phones/laptops) during the test

- 2. **One piece of Coursework: — 25%**
  - Due date: **5PM on May 20**
  - A mini project using R for fitting generalised linear models
  - Project details to be handed out later

- 3. **Written final assessment: — 50%**
  - Time: end of semester exam period
  - Based on both linear (mixed) and generalised linear models

Contact and Office Hour

- **Module mailbox:**
  - math3091@soton.ac.uk.
  - All queries should be sent to this email address only.
  - We both will monitor this email address.
  - We aim to give you a prompt reply.
  - E-Mail sent to our university email addresses may receive a delayed reply.

- **Office Hour:**
  - Monday 16:00-17:00 (during Weeks 1-5)
  - Friday 14:30-15:30 (TBD, during Weeks 6-10)
  - No prior appointment is needed
  - Bb Collaborate, or see me in person at 9001/B54 (during weeks 1-5).

The Module Blackboard Site

- **Module Information:**
  - provides the information presented above

- **Diary (log book)**
  - Organised by week, this is where you will find all the lecture videos, slides, R work sheets for computer labs, theory exercise sheets for the problem classes
  - After the end of the live class we will also publish recordings and exercise solutions.
  - This page will also detail what activities are planned for the week ahead at any particular time

- **Course Content**
  - Lecture note booklet
  - R data sets
  - Past papers and solutions
  - Any additional material

What we will be learning and why is it exciting?

- We will be modelling realistic real life datasets with statistical models, which help us to
  - **understand** the random process by which observed data have been generated.
  - **make predictions and decisions** based on our inferences concerning the process.
  - **assess uncertainties** in the inference we make

- We will extend and develop new methods for model fitting
  - Linear (mixed) models extended from MATH2010
  - Generalised linear models for count data

- Below are some example datasets we will analyse:

1.1 Dataset Examples

`nitric`: Nitric acid

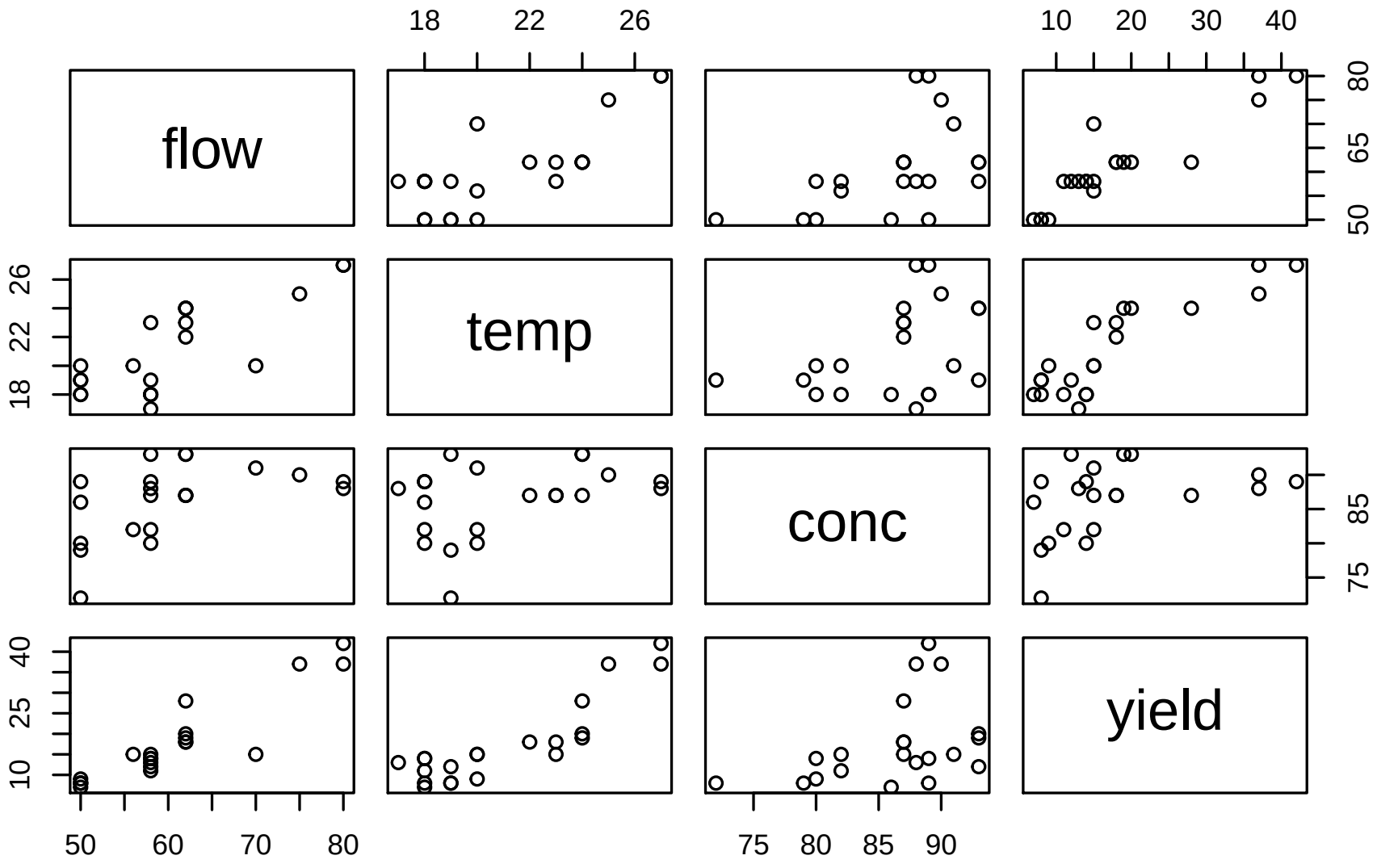
This data set relates to 21 successive days of operation of a plant oxidising ammonia to nitric acid.

The response `yield` is ten times the percentage of ingoing ammonia that is lost as unabsorbed nitric acid (an indirect measure of the yield).

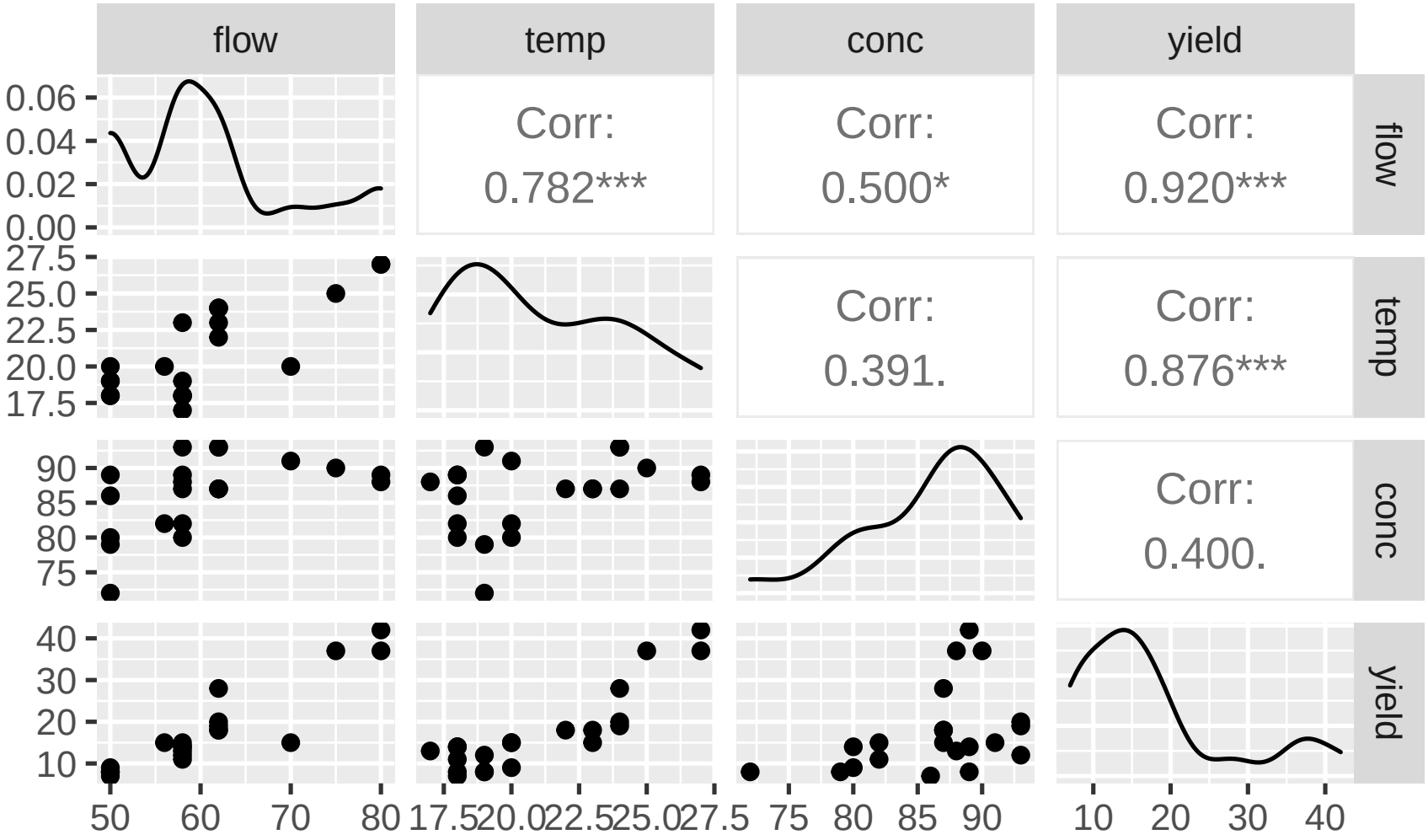
The aim here is to study how the yield depends on:

- flow of air to the plant (`flow`),
- temperature of the cooling water entering the absorption tower (`temp`)
- concentration of nitric acid in the absorbing liquid (`conc`).

Plotting the `nitric` data (pairs plot from Math2010)



Plotting the `nitric` data (Learning more from ggplot)

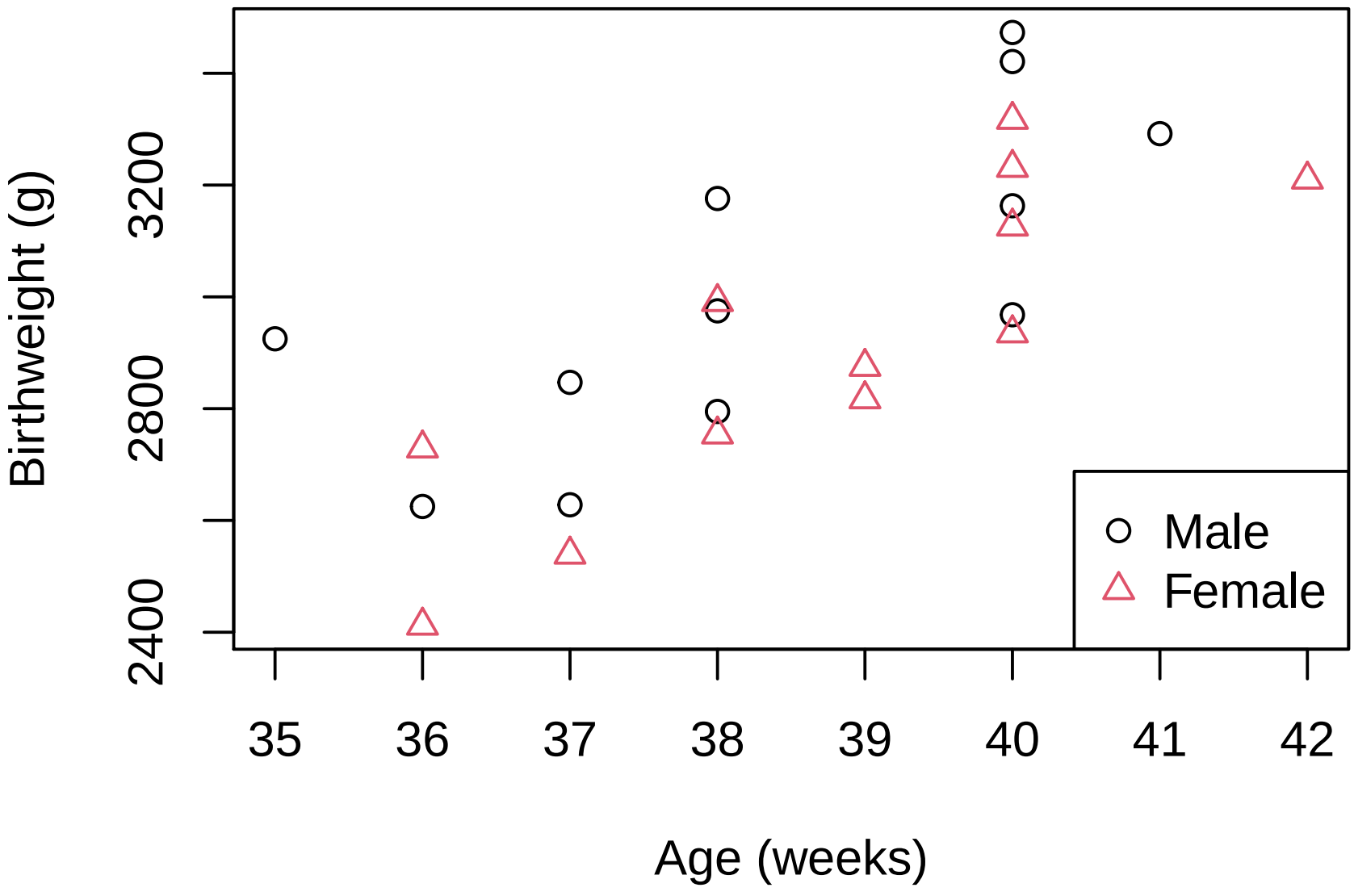


`birth`: Weight of newborn babies

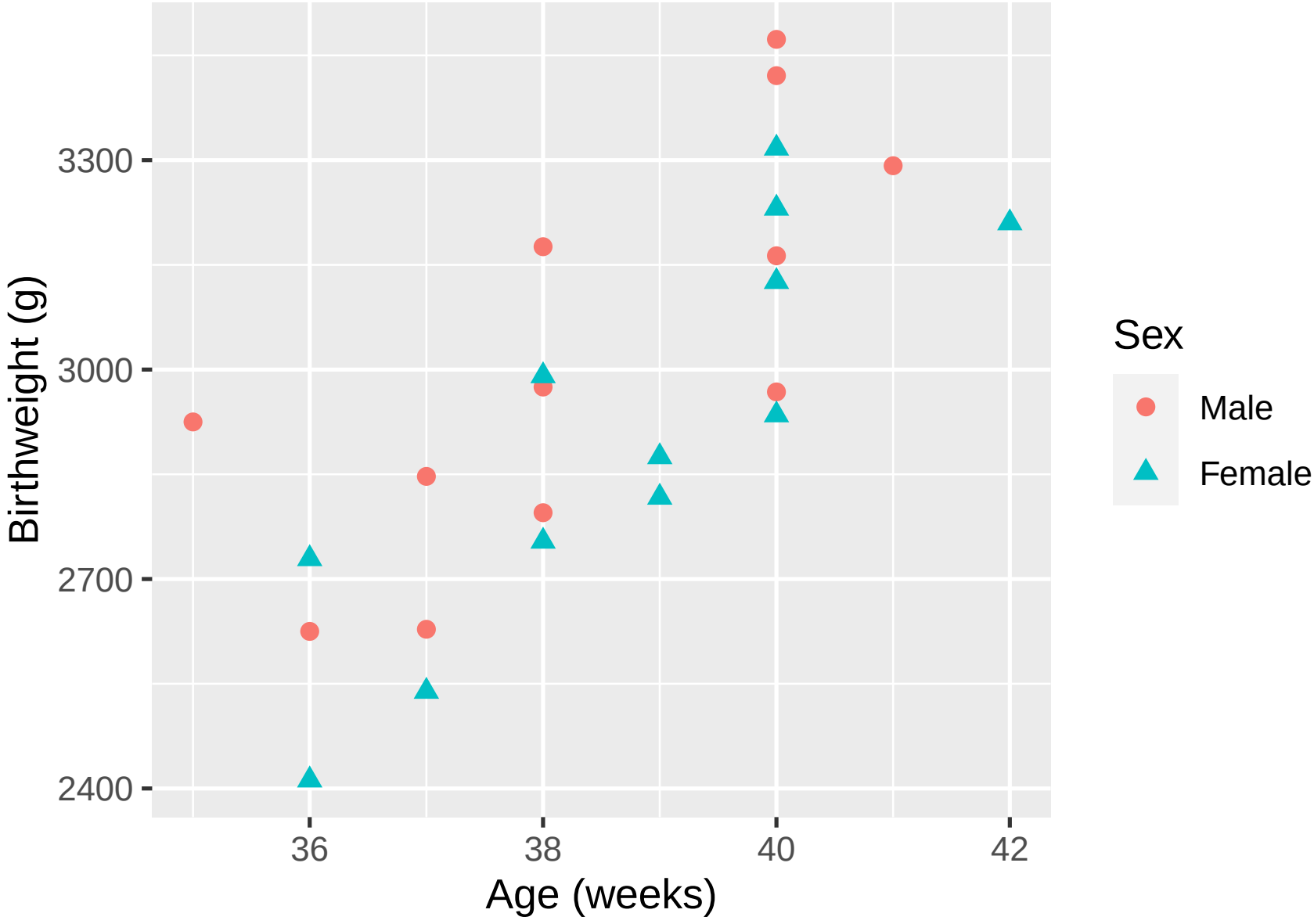
This data set contains weights of 24 newborn babies. There are two explanatory variables, sex (`Sex`) and gestational age in weeks (`Age`) together with the response variable, birth weight in grams (`Weight`).

The aim here is to study how birth weight depends on sex and gestational age.

Plotting the `birth` data (basic scatter plot)



Plotting the `birth` data (from ggplot 2)

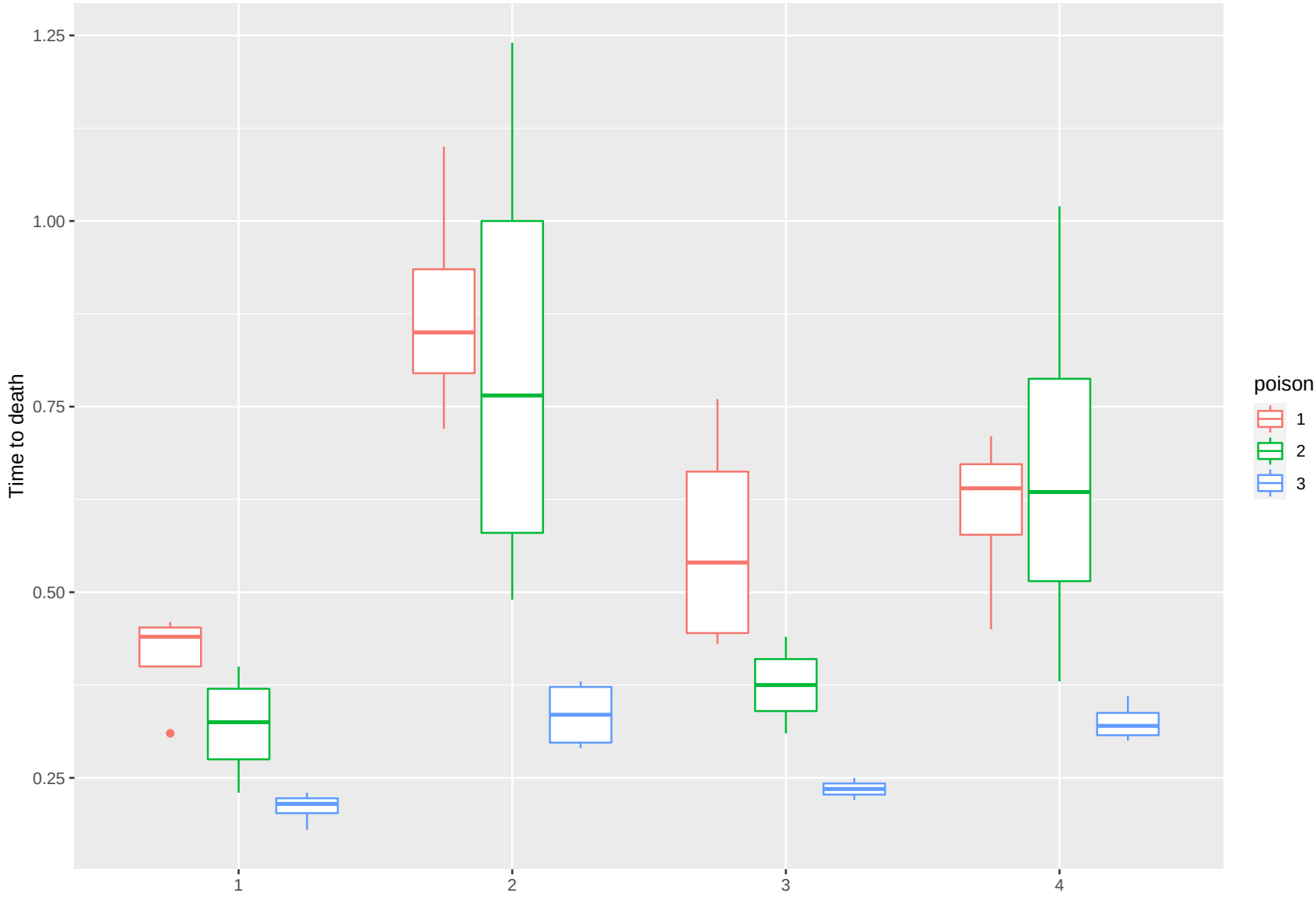


`survival`: Time to death

This data set contains survival times in 10 hour units (`time`) of 48 rats each allocated to one of 12 combinations of 3 poisons (`poison`) and 4 treatments (`treatment`).

The aim here is to study how survival time depends on the poison and the treatment, and to determine whether there is an interaction between these two categorical variables.

Plotting the `survival` data

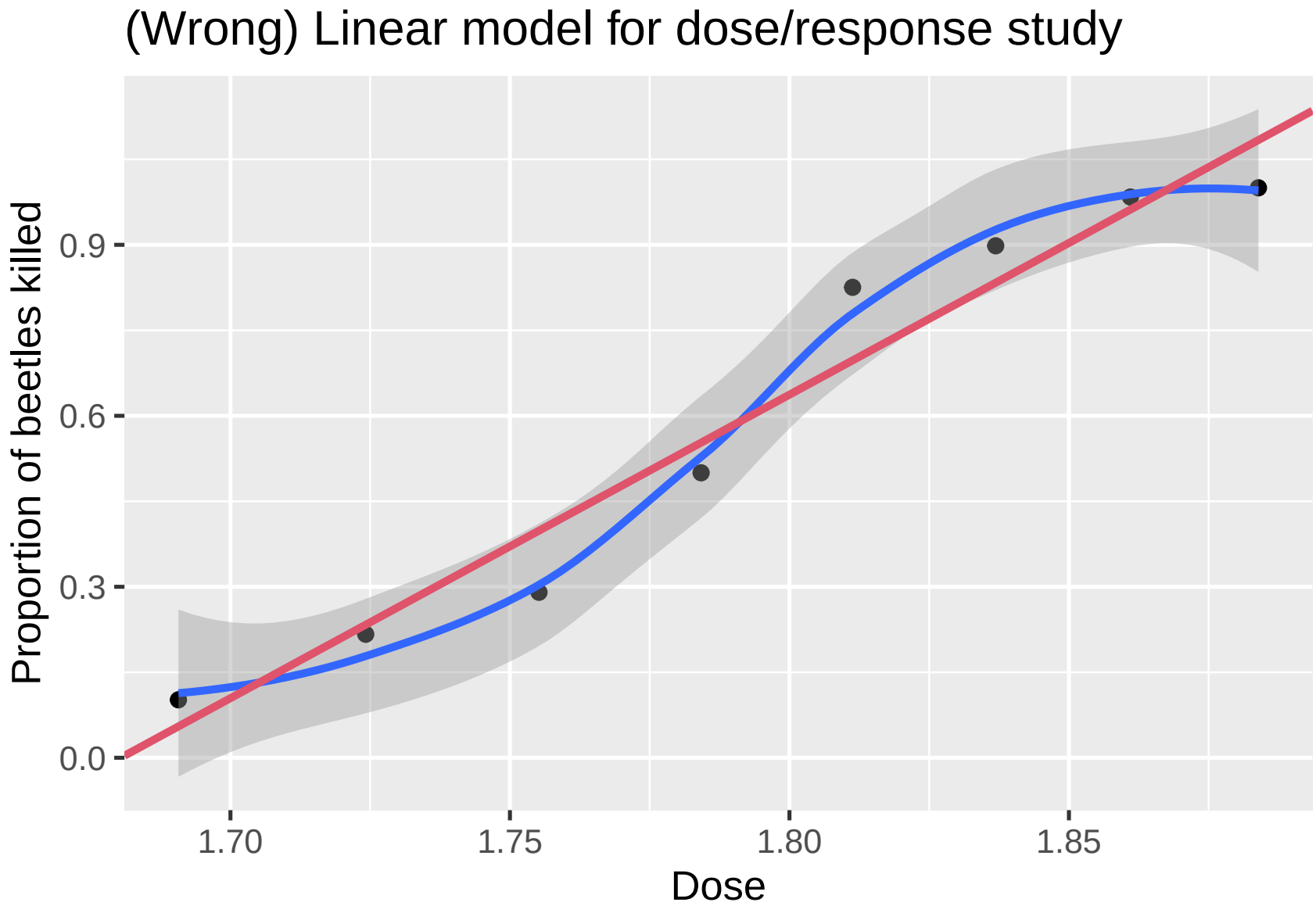


beetle: Dose response study on insecticide

This data set represents the number of beetles exposed (exposed) and number killed (killed) in eight groups exposed to different doses (dose) of a particular insecticide. Interest is focussed on how mortality is related to dose.

```
##      dose exposed killed
## 1 1.6907      59      6
## 2 1.7242      60     13
## 3 1.7552      62     18
## 4 1.7842      56     28
## 5 1.8113      63     52
## 6 1.8369      59     53
```

The proportion of beetles killed by dose of insecticide



`shuttle`: Challenger disaster

This data set concerns the 23 space shuttle flights before the Challenger disaster.

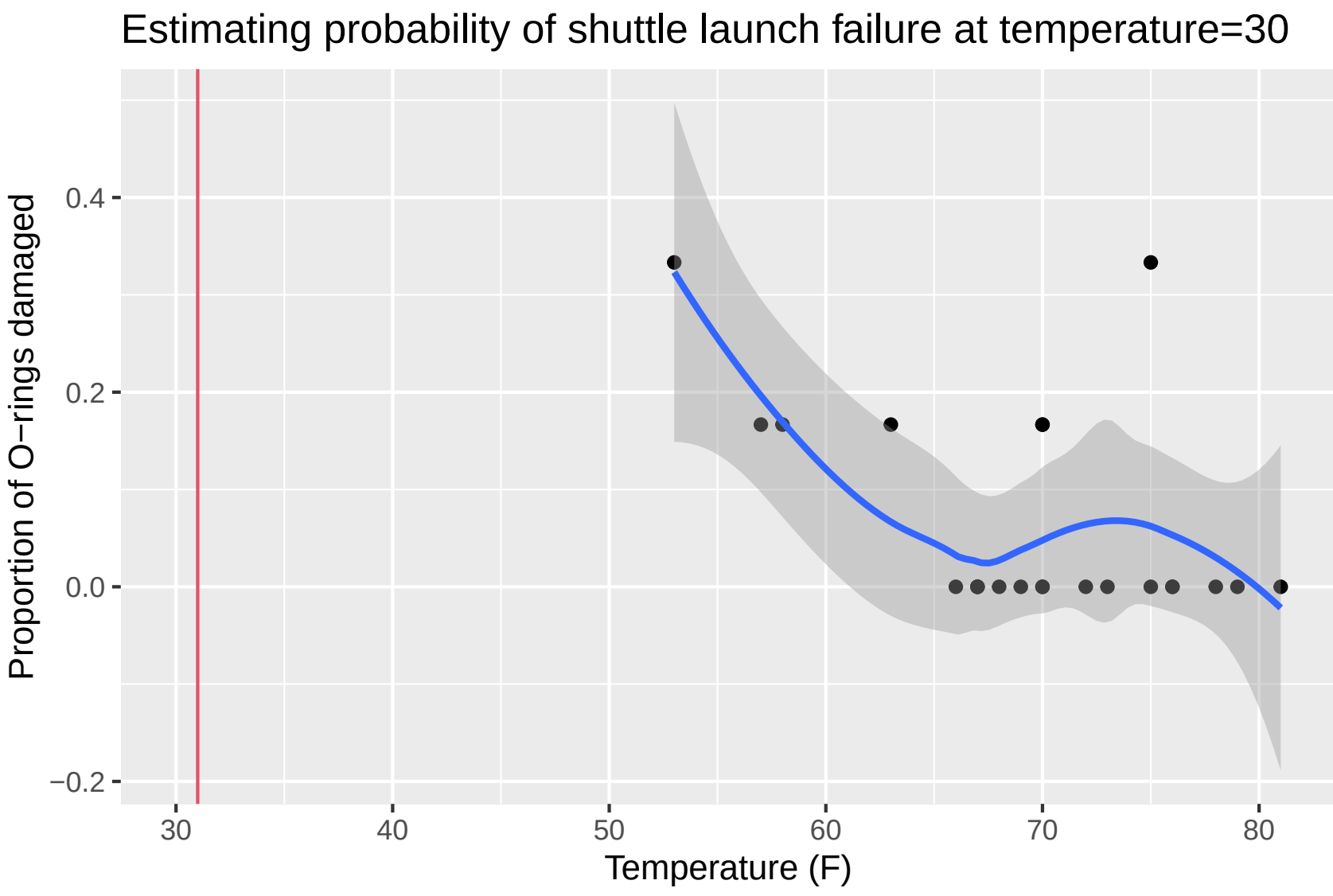
The disaster is thought to have been caused by the failure of a number of O-rings, of which there were six in total.

The data consist of variables including:

- the number of damaged O-rings for each pre-Challenger flight
- the launch temperature in degrees Fahrenheit
- the pressure at which the pre-launch test of O-ring leakage was carried out
- the name of the orbiter

The Challenger launch temperature on 20th January 1986 was 31F. The aim is to predict the probability of O-ring damage at the Challenger launch.

The proportion of O-rings damaged by launch temperature



Other data sets analysed

- 'heart': Estimating treatment effects for heart attack patients
  - Which factors influence the most patient survival rate after heart attack?
- 'accident': Number of road accidents in Cambridge
  - Which of the two roads is more dangerous?
  - How does time of day affect road accident?
- 'job': Job satisfaction and income
  - Are people with higher income more satisfied with their jobs?

Modelling pandemic data?

Covid-19 deaths in different local authorities

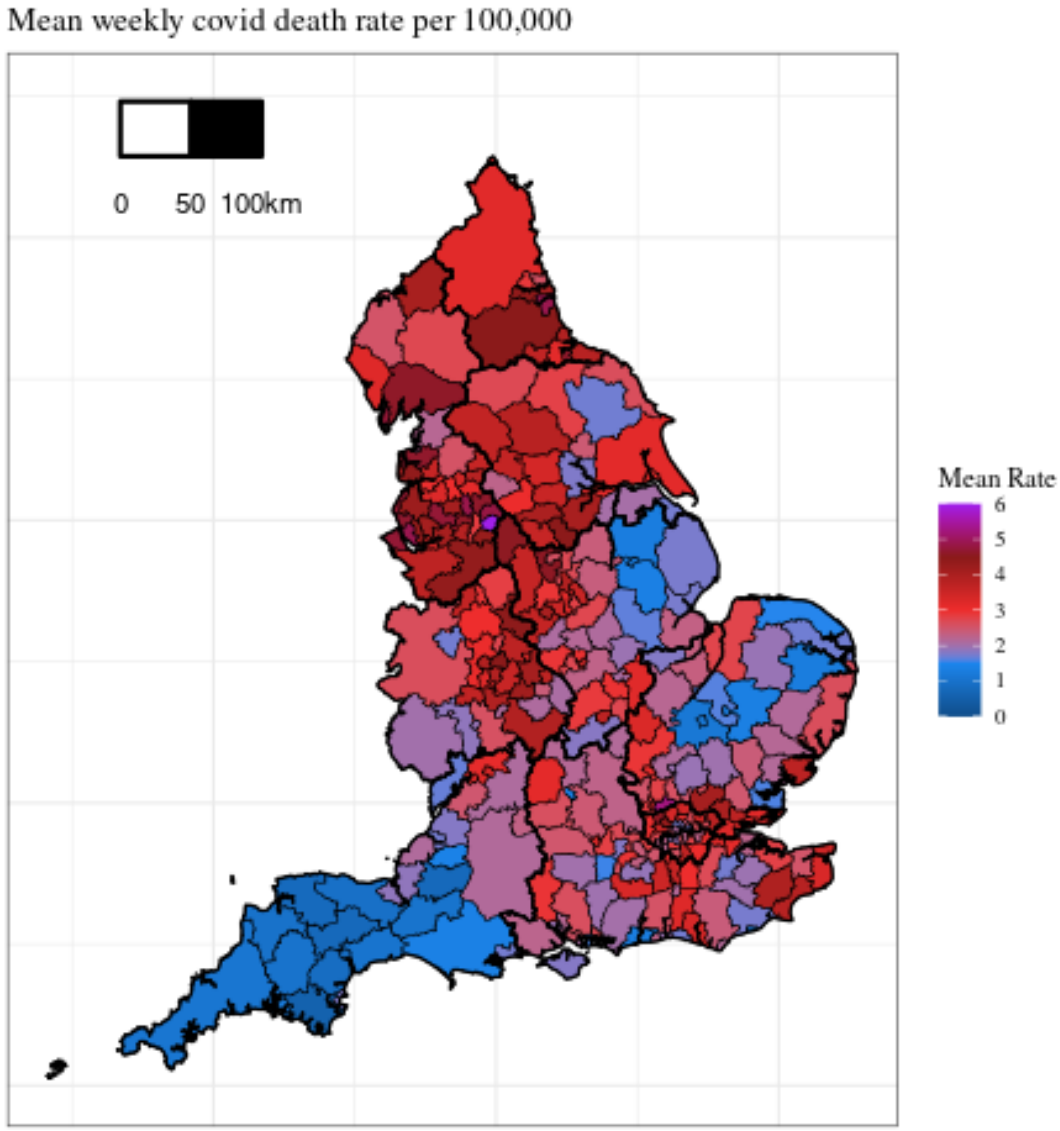


Figure 1.1

1.2 Elements of statistical modelling

Statistical modelling

Statistical models help us to

- **understand** the random process by which the observed data have been generated (data generating mechanism).
- make **predictions** and **decisions** based on our **inferences** concerning the process.

Statistical modelling

Statistical models help us to

- **understand** the random process by which the observed data have been generated (data generating mechanism).
- make **predictions** and **decisions** based on our **inferences** concerning the process.

We describe random processes using **probabilistic models**. Given a model, statistical inference may involve **estimating** any unspecified features of the model, **comparing** competing models, assessing the **appropriateness** of a model; and report **uncertainties** associated with the conclusions we draw, all in the light of observed data.

- Above are methods of mathematically approximating the world
- People from CS background may also call above as *"Machine Learning"*

Requirements of a statistical model

- **Plausibility**: the model should be reasonable, given background (scientific) knowledge of the phenomenon being studied.

- **Parsimony**: we should avoid the model being more complicated than it needs to be.

Occam's razor: entities should not be multiplied beyond necessity

- **Goodness of fit**: the model should fit the data well.

All models are wrong
but some are useful

George E.P. Box

1.3 (Linear) Regression Models

Regression Models

In practical applications, we often distinguish between a *response* variable and a group of *explanatory* variables.

The aim is to determine the pattern of dependence of the response variable on the explanatory variables. A regression model has the general form

$$\textit{response} = \textit{function(structure and randomness)}$$

- *structure* describes how the response depends on the explanatory variables
- *randomness* describes the probability distribution of the response.

Linear models

In MATH2010, you studied linear models, where the response is assumed to follow *normal* distribution, and its mean depends on the explanatory variables.

$$Y_i = X_i\beta + \varepsilon_i, \quad i = 1, \ldots, n$$

Linear models

In MATH2010, you studied linear models, where the response is assumed to follow *normal* distribution, and its mean depends on the explanatory variables.

$$Y_i = X_i\beta + \varepsilon_i, \quad i = 1, \ldots, n$$

Is the assumption always a reasonable model?

Distributions quiz 1

Which distribution could be used to model the number of students enrolled in a course who attend a given lecture?

- Bernoulli
- Binomial
- Exponential
- Gamma
- Geometric
- Normal
- Poisson
- None of the above

Distributions quiz 1

Which distribution could be used to model the number of students enrolled in a course who attend a given lecture?

- Bernoulli
- Binomial
- Exponential
- Gamma
- Geometric
- Normal
- Poisson
- None of the above

We might want to model how that number of students varies according to some explanatory variables, such as time of day, day of the week, week of the semester, and so on.

That requires a more general regression model.

Go beyond the linear model

- linear mixed model
- generalised linear model
- nonparametric regression
- quantile regression
- Cox proportional-hazards model
- ...

Conclusion

- We have seen some examples of situations in which it is not sensible to use the linear model. We will see some more examples in this module.
- Once we have written down the model, we want to **estimate** the parameters of the model, express **uncertainty** in estimates, make **predictions** and **compare** candidate models.
- We learned many models beyond linear regression
- We can do all these using the **likelihood** based methods.