Today’s agenda

▶ Introductory Comments
  ▶ House keeping
  ▶ Computer details
  ▶ Plan of the course

▶ Statistical modelling: an overview
  ▶ Our analysis strategy
  ▶ Goals for the course

▶ Role of graphics

▶ Data cleaning
Housekeeping

- **Contact details**

<table>
<thead>
<tr>
<th>Office</th>
<th>Email</th>
</tr>
</thead>
<tbody>
<tr>
<td>Steffen Klaere</td>
<td>303.219</td>
</tr>
<tr>
<td>Alan Lee</td>
<td>303S.265</td>
</tr>
</tbody>
</table>

- **Course material on course website**


- **Office hours**

  | Steffen Klaere       | 10:00–12:00, Thu |
  | Alan Lee             | 10:30–12:00, Tue and Thu |

- **Class Rep?**
STUDENTS HAVE A RIGHT TO BE HEARD AT THIS UNIVERSITY!
BECOME A CLASS REP AND BE THE VOICE OF THEIR CHOICE.

WAVE SUPPORT FOR CLASS REPS INCLUDES:
TRAINING, FUNDING FOR CLASS PARTIES, CLASS REP CERTIFICATES
AND ONGOING ADVICE.

CONTACT US AT:
309 0789 EXT 202 OR VIA EMAIL: WAVE@AUCKLAND.AC.NZ
OR VISIT US AT THE STUDENT AMENITIES BUILDING, LEVEL 2.
Welcome to the STATS 330/762 Homepage

This web page is designed to keep you informed about the course. It contains copies of the assignments (including the data required) and model answers, together with tests, assignments and exams from past years. To supply feedback on any aspect of the course, including these web pages, email Steffen or Alan.

Taught by:

Steffen Klaere           Alan Lee

Contact details:

Steffen: Stop by my office, Room 303-219, call me on 373-7599 Extn 85237 or send me an email.
Alan: Stop by my office, Room 303S-265, call me on 373-7599 Extn 88749 or send me an email.
Assignments

- There will be five assignments
  - 20% of total grade

- The due dates are listed in the course information
  - Assignment 1 is due August 7

- No paper will be issued in class, download assignments and data from the web or cecil.
Tutorials

- These will cover computing details
- Held Fridays in tutorial lab, 303S-G75
  - Tutorial 1: 10:00–11:00
  - Tutorial 2: 13:00–14:00
  - Tutorial 3: 16:00–17:00
- Start second week of semester
Computer details

- All analyses will be done using RStudio.

- Homework needs to be typed, copy and paste R-output into results.
  
  **WORD:** remember to use Courier or similar monospaced fonts.

  **LaTeX:** Use verbatim environment for simple copy and paste or sweave package to combine R and LaTeX.

- Use home/laptop computer or lab (303S).

- See web page for info on downloading R330 package.

Course plan

We aim for 33 lectures divided into chapters as follows

Chapter 1: Introduction (1 lecture, this one!)

Chapter 2: Graphics (3 lectures)

Chapter 3: Multiple Regression models (12 lectures)

Chapter 4: Factors (4 lectures)

Chapter 5: Models for categorical and count responses (12 lectures)

Revision: 1 lecture
Course book

- This covers most of the material we will discuss in the lectures
- Has five chapters corresponding to the division on the previous slide
- Online version available on the course web site
- Offline version available from the Statistics Desk, 303, 3rd floor.
Overview of Statistical Modelling

- Statistical models summarise relationships between variables.
- Regression models focus on one response variable and how its distribution can be modelled by one or more explanatory variables.
Example: Cute babies

**BPD:** Biparietal Diameter

For an unborn baby, depends on several factors, including Gestational Age.
Ultrasound image
Points to take into account

- Not all babies of the same age have the same BPD.
- In general, the older the baby, the bigger the BPD.
Model relating BPD and GA

Mean BPD for 30 weeks = $\alpha + \beta \cdot 30$

Mean BPD for 40 weeks = $\alpha + \beta \cdot 40$

Mean BPD for GA weeks = $\alpha + \beta \cdot \text{GA}$
Regression model features

- Model specifies the distribution of the response.
- Also says how the distribution of the response is affected by the covariates (explanatory variables).
- In this case the covariate GA determines the mean response:

\[
\text{mean BPD} = a + b \cdot \text{GA},
\]

i.e. a straight-line relationship.
In general

- **Response** has a distribution depending on the **explanatory** variables

- Mean of the distribution given by some function of the explanatory variables

- Need to
  - describe the function
  - describe the variability

- Balance accuracy with simplicity
Our analysis strategy

- Explore the data using graphics and summary statistics
- Construct a useful model (trial and error)
- Use the model to gain knowledge about the system under study (How big should a 40 week old baby’s BPD be?)
- Communicate findings (be able to write a report!!!)
Goals for 330

- Get more practice in exploring data
- Expand knowledge of regression
- Get better at fitting models
- Improve your diagnostic skills (how to recognise when a model does not describe the data properly?)
- Improve your interpretation and communication skills, in a more flexible way.
Role of graphics

Data cleaning: Are there gross errors, outliers, special codings (e.g., 999 for missing), missing data, absolute rubbish in the data?

Exploratory analysis: What sort of model might be appropriate?

Diagnostics: Having tentatively selected a model, is it any good? (residual plots, etc.)
Exploratory analysis for BPD data
Data Cleaning

- All real-life data sets are likely to contain errors
- These will usually be revealed by suitable plots, such as the one on the previous slide
- Before starting any analysis, data should always be carefully checked
- To encourage this practice, we will occasionally introduce errors into the assignment data supplied on the web
Data Cleaning

It is your responsibility to check all data supplied on the web, against the assignment sheets. Failure to do so will cost marks.
Data Cleaning

It is your responsibility to check all data supplied on the web, against the assignment sheets. Failure to do so will cost marks.

YOU HAVE BEEN WARNED