Session3:
Regression analyses

Dr Nicola Foster
UCL Respiratory
nicola.foster@ucl.ac.uk

All code for this session illustration is on
GitHub: nicolacfoster

Outline of this session

Building on our last session about statistical inference, we now look at regression modelling – statistical models for inference or predicting an outcome variable from a set of predictors.

1. What is regression analysis?
   - Linear regression
   - Logistic regression
   - Non-linear models
2. Constructing regression models
   - Interaction terms
   - Diagnostic tests
   - Model assumptions
3. Fitting regression models to data
   - Visualising model fits
   - Understanding residuals
   - Comparing models
4. Interpreting outputs
5. Briefly introduce multi-level regression models

What is regression analysis

Regression analyses are analyses using statistical models to predict an outcome variable from a set of predictors. We can also use statistical models to explain the relationship between two (or more) variables (inferential).

When deciding which statistical model to use, you need to consider:

- Outcome variable that you are modelling (continuous, categorical, ordinal or count).
- Assess whether the relationship between the dependent and independent variables are linear or non-linear (model fitting process).

Using the R datasets package/ Kaggle, I'll work through an example of regression analysis and the interpretation.

```
library(datasets)

# identifies datasets to use for analyses
# library(help="datasets")
```

Case study: Salt & Temperature

Linear regression analysis: predicting/ explaining a continuous value outcome, and assumes that the relationship between the outcome and predictors can be represented by a straight line, and written by the following equation.

$$y = mx + c$$

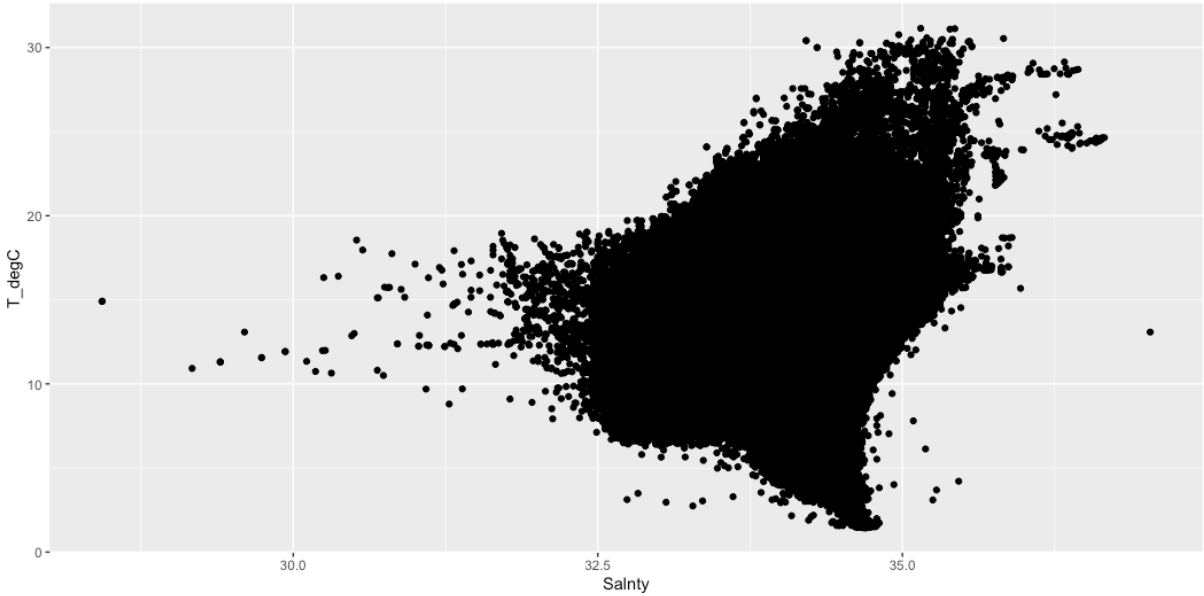
Using a practice dataset, we'll answer the following questions:

- Is there a relationship between salt concentration & water temperature?
- Can you predict water temperature based on salt concentration?

Plotting the relationship between water salinity & temperature

- Load and view the data
- Review the data documentation
- Plot the relationship between salinity and temperature, using a dot plot
- Looking at the plot, does this relationship look linear to you?
- The relationship looks roughly linear, with some caveats, let's evaluate this some more.

```
library(ggplot2)

data1 <- read.csv("bottle.csv")
attach(data1)

ggplot(data1, aes(Salnty, T_degC), panel
= panel.smooth()) + geom_point()
```



Assessing correlation

- Correlation coefficient: a statistical measure that quantifies the strength and direction of the linear relationship between two variables.
  - A perfect negative linear relationship, would be -1, while a perfect positive linear relationship would be +1. A 0 indicates that there is no linear correlation between two variables.
  - Moderate correlation: 0.5-0.7; while a high correlation is between 0.7 and 0.9.
- In our example, the correlation coefficient is -0.51. Suggesting a moderate negative relationship or as salt concentration increases, water temperature decreases.
- We want to understand the non-linearity/ where the relationship deviates from linear better.

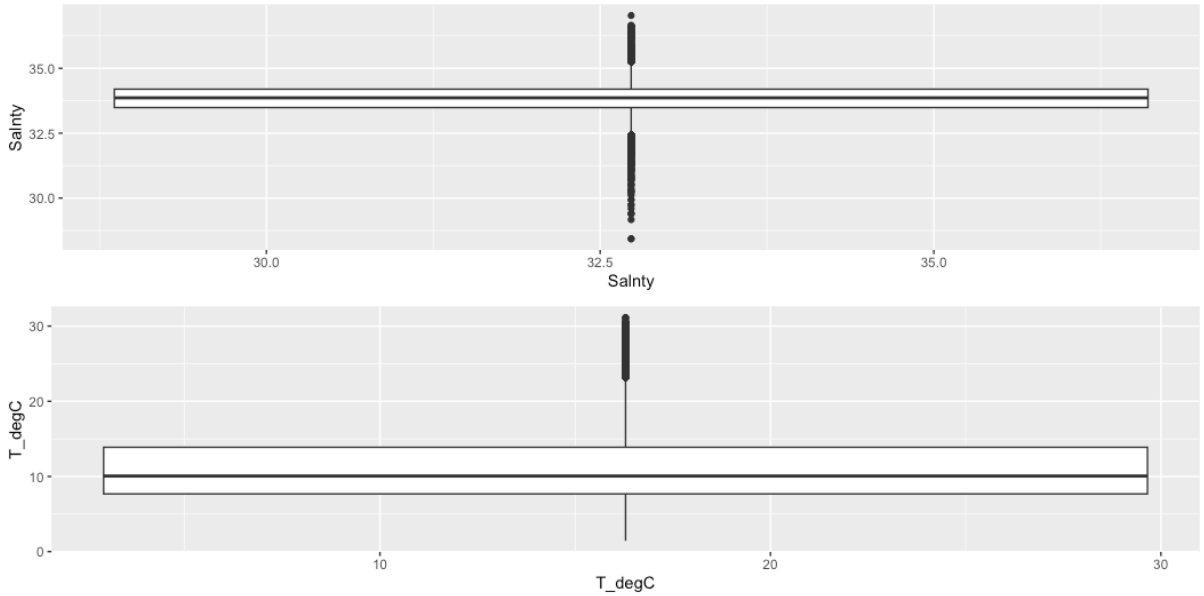
```
# Find the correlation coefficient
cor.test(Salnty, T_degC)
```

```
    Pearson's product-moment correlation

data:  Salnty and T_degC
t = -528.33, df = 814245, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.5068817 -0.5036466
sample estimates:
       cor
-0.505266
```

Consider data outliers

- A reason for non-linearity may be outliers
- Boxplots of the variables suggests the presence of outliers

```
library(gridExtra)

# Check for outliers by producing a box plot showing the distribution of the variables
plot1 <- ggplot(data1)+
  geom_boxplot(aes(Salnty, Salnty))

plot2 <- ggplot(data1)+
  geom_boxplot(aes(T_degC, T_degC))

grid.arrange(plot1, plot2)
```



Inferential versus predictive modelling

The results so far suggests that the relationship is fairly linear, but that the data are not fully explained by only the one variable we've included so far.

If you are using the model to predict outcomes or if you are working on a machine learning analysis, you may want to split your data (randomly) into a training and test dataset before fitting the model.

Now, create a model that predicts water temperature based on salinity.

Specifying the model

```
# Create your linear regression models
watertemp <- lm(T_degC~Salnty, data=data1)
watertemp
```

```
Call:
lm(formula = T_degC ~ Salnty, data = data1)

Coefficients:
(Intercept)       Salnty
    167.351       -4.624
```

Review residual standard errors

Review the residual standard error (RSE) of each model.

```
sigma(watertemp)
```

```
[1] 3.645963
```

In other words, the true water temperature, will be off by 3.65 degrees Celsius from what the model predicted. Let's assess the model fit, more formally. For linear regression models, we can do this by reviewing the R squared value of the model `watertemp`.

Review model fit (R squared)

The R squared value expressed the proportion of variance explained in the model, we could therefore say that this model does not fully explain the data. Only 26% of the variance in salinity explains the water temperature.

```
summary(watertemp)$r.squared
[1] 0.2552937
```

Review full model output

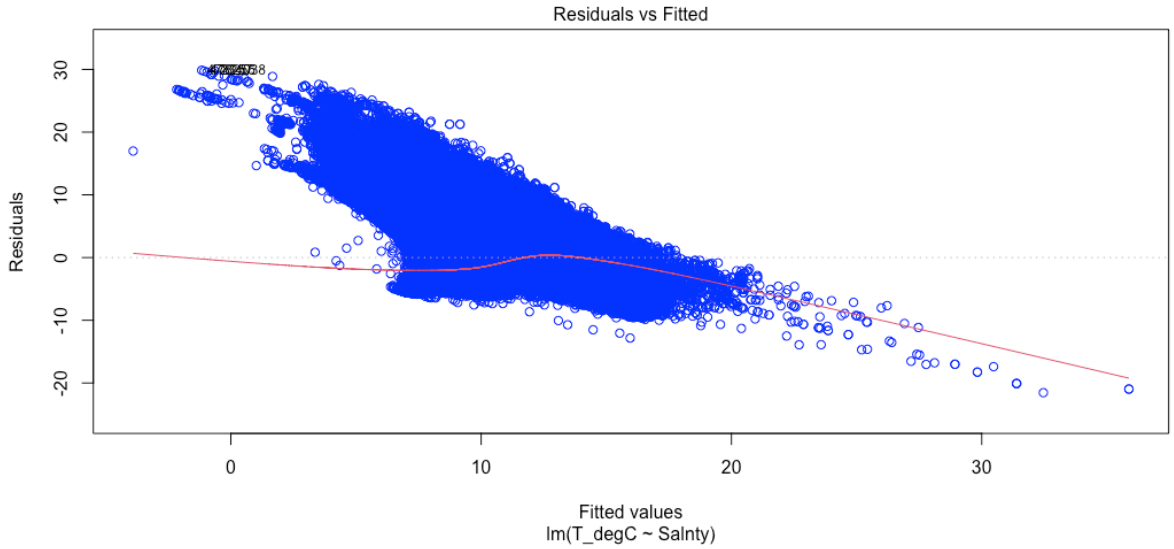
Look at the full summary of the model output:

```
summary(watertemp)

Call:
lm(formula = T_degC ~ Salnty, data = data1)

Residuals:
    Min      1Q  Median      3Q     Max
-21.542  -2.298  -1.033   1.496  29.866

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 167.350630   0.296226   564.9   <2e-16 ***
Salnty       -4.624236   0.008753  -528.3   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.646 on 814245 degrees of freedom
  (50616 observations deleted due to missingness)
Multiple R-squared:  0.2553,    Adjusted R-squared:  0.2553
F-statistic: 2.791e+05 on 1 and 814245 DF,  p-value: < 2.2e-16
```

Review model fit (residuals versus fitted)

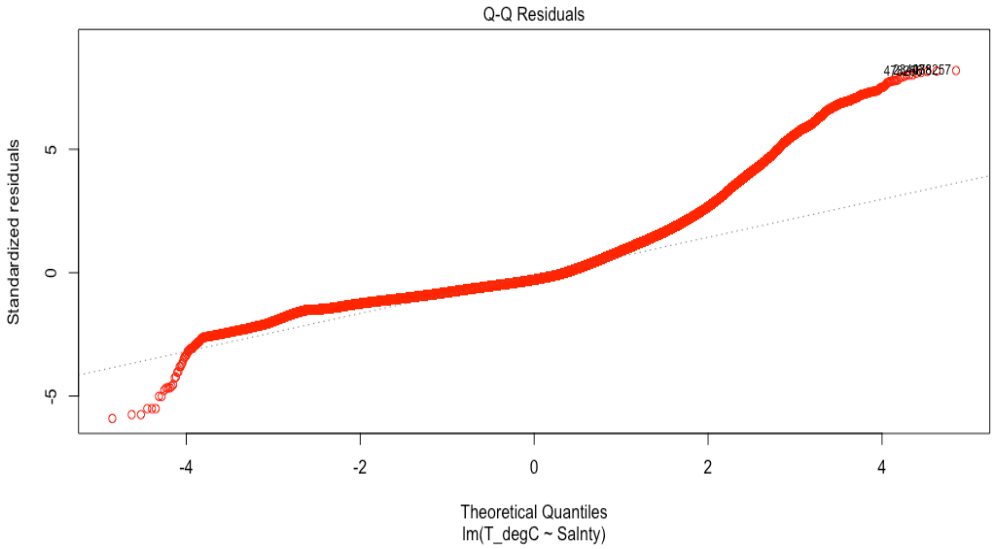
Now we plot the residuals to further assess model fit.

```
plot(watertemp, which=1,
col=c("blue")) #residuals versus
fitted plot
```



Review model fit (Quantile plot)

- Another way to view fit, to plot the quantile plot - to assess whether the residuals are normally distributed
- If residuals follow close to a straight line on this plot ,- normal distribution.
- Here, we notice initial good alignment but moves off the line considerably so suggests not an ideal fit - which confirms our review of the R2 values for the model.
- Are there other variables that may help us to explain this relationship?




```
plot(watertemp, which=2, col=c("red"))
```

Multivariate regression

- A statistical technique used to model/ explain the relationship between multiple dependent variables and one/ more independent variables.
- Predicting the values of the dependent variables based on the values of the independent variables.
- Would a multivariate regression model fit better? Are there other variables included in the dataset that would help us to explain water temperature better?

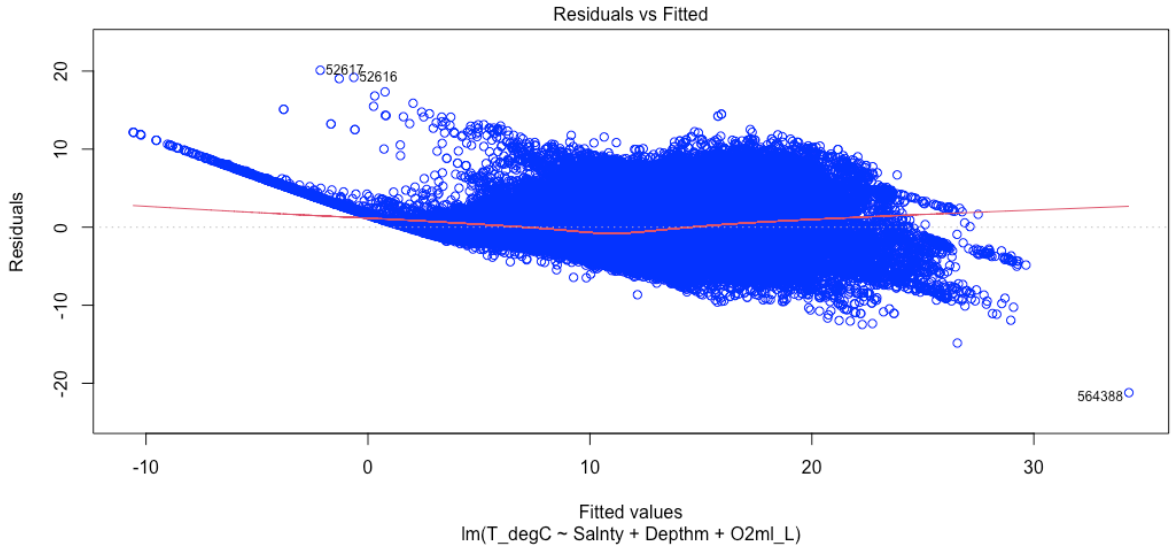
Specifying the model

- Process of model specification may include a **directed acyclic graph**, to specify your assumptions
- Or may be based on your available data – data driven approach to inference (machine learning).

```
# Create your linear regression models
watertemp2 <- lm(T_degC~Salnty + Depthm + O2ml_L, data=data1)
watertemp2

Call:
lm(formula = T_degC ~ Salnty + Depthm + O2ml_L, data = data1)

Coefficients:
(Intercept)       Salnty       Depthm       O2ml_L
 -1.683e+02    5.115e+00   -5.012e-03    2.118e+00
summary(watertemp2)

Call:
lm(formula = T_degC ~ Salnty + Depthm + O2ml_L, data = data1)

Residuals:
     Min       1Q   Median       3Q      Max
-21.2022  -1.1723  -0.1595   0.8195  20.1096

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.683e+02  3.200e-01  -525.8   <2e-16 ***
Salnty       5.115e+00  9.309e-03   549.5   <2e-16 ***
Depthm      -5.012e-03  9.647e-06  -519.5   <2e-16 ***
O2ml_L       2.118e+00  2.117e-03  1000.4   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.944 on 661485 degrees of freedom
  (203374 observations deleted due to missingness)
Multiple R-squared:  0.7884,    Adjusted R-squared:  0.7884
F-statistic: 8.213e+05 on 3 and 661485 DF,  p-value: < 2.2e-16
```

Review model fit

```
plot(watertemp2, which=1,
col=c("blue")) #residuals versus
fitted plot
```



Adding confidence intervals

```
confint(watertemp2)
                    2.5 %        97.5 %
(Intercept) -1.689023e+02 -1.676479e+02
Salnty       5.096880e+00  5.133373e+00
Depthm      -5.030798e-03 -4.992981e-03
O2ml_L       2.113832e+00  2.122131e+00
```

Interpreting the coefficient

- For linear regression, we interpret the regression coefficient as the difference between two marginal means (“Mean Difference”), when you have chosen values of X that are one unit apart.
- For the coefficients of a logistic regression analysis, exponentiate the coefficient (raising them to the power ‘e’) to make it easier to interpret them.
- The exponentiated coefficient is interpreted as the Odds Ratio, which means that the odds of the event occurring is increased by a factor of 1.XX or by XX%

```
# dichomotised the outcome variable
data1$T_degCbin <- as.numeric(data1$T_degC >= 3)
table(data1$T_degCbin, useNA="ifany")

     0      1   <NA>
  7469 846431  10963
```

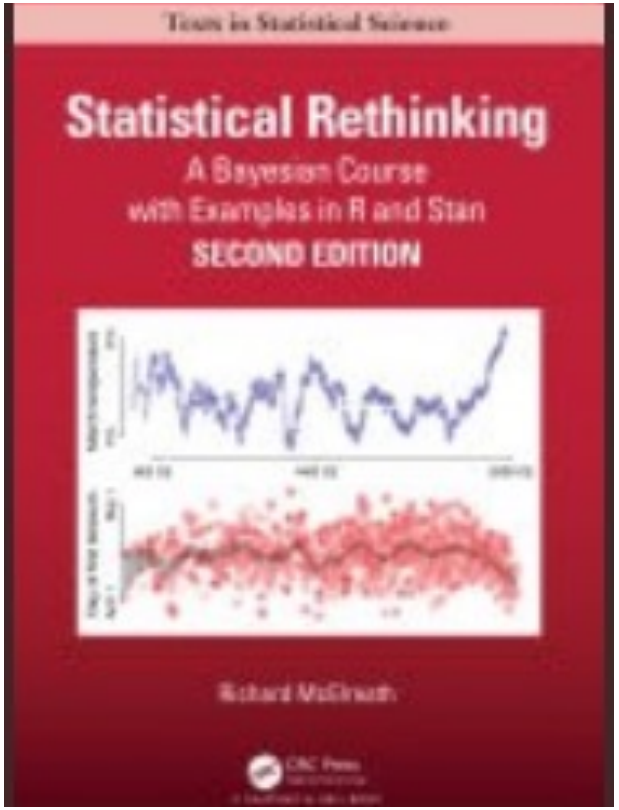
```
watertemp3 <- glm(as.factor(T_degCbin)~Salnty + Depthm + O2ml_L,
family=binomial(link="logit"), data=data1)
watertemp3

Call:  glm(formula = as.factor(T_degCbin) ~ Salnty + Depthm + O2ml_L,
    family = binomial(link = "logit"), data = data1)

Coefficients:
(Intercept)       Salnty       Depthm       O2ml_L
 -125.46001      4.30949     -0.01656     -0.03057

Degrees of Freedom: 661488 Total (i.e. Null);  661485 Residual
  (203374 observations deleted due to missingness)
Null Deviance:      68120
Residual Deviance: 3651     AIC: 3659
summary(watertemp3)

Call:
glm(formula = as.factor(T_degCbin) ~ Salnty + Depthm + O2ml_L,
    family = binomial(link = "logit"), data = data1)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.255e+02  1.222e+01 -10.266   <2e-16 ***
Salnty       4.309e+00  3.613e-01  11.927   <2e-16 ***
Depthm      -1.656e-02  4.136e-04 -40.042   <2e-16 ***
O2ml_L      -3.057e-02  1.683e-01  -0.182    0.856
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 68124.6  on 661488  degrees of freedom
Residual deviance:  3650.6  on 661485  degrees of freedom
  (203374 observations deleted due to missingness)
AIC: 3658.6

Number of Fisher Scoring iterations: 15
exp(coefficients(watertemp3))
  (Intercept)       Salnty       Depthm       O2ml_L
3.261431e-55 7.440216e+01 9.835733e-01 9.698886e-01
```

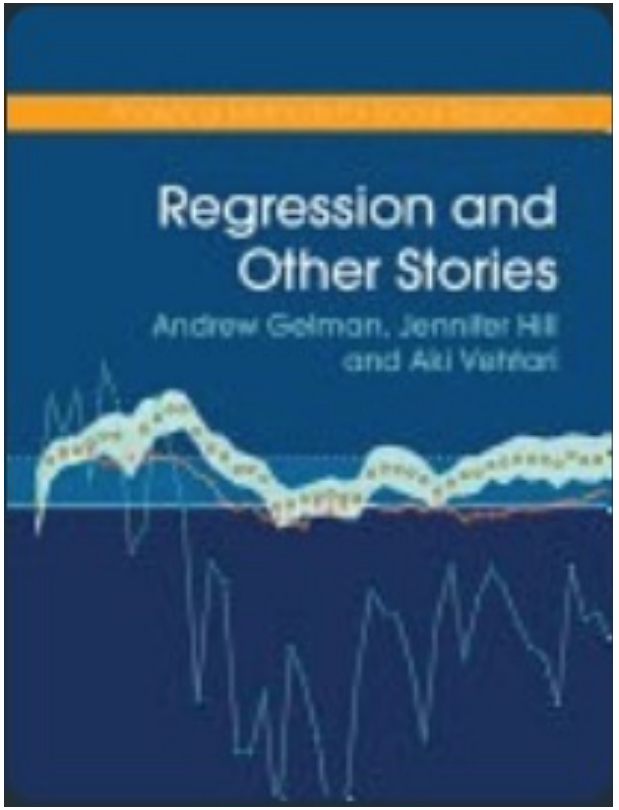
Other approaches/ steps to consider

Depending on the model fit to the data, you may want to consider the following:

- Assess missing values and they may affect your analysis. Would you consider imputing missing values? If imputing values, you have to report both your complete case analysis (dropped missing observations) as well as the analyses of data that includes imputation.
- Non-linear models, where you use types of piecewise polynomial/ (cubic) spline curve that fits a series of data points.
- Assess whether there is clustering in your data, this could be by (i) geographical location, (ii) time step, or (iii) within groups.

Further reading





Schedule

Sep	Programming languages	Data science introduction session with a focus on setting up to code in R, showing people how to import datasets, summarise data descriptively and where to find help.
Feb	Statistical thinking	Introduction to statistical thinking, types of data and how data types influence the methods we use, data management while doing the analysis, and data processing tips.
Apr	Regression analyses	Introduction to regression analyses, identifying covariates for your models, correlation and confounding.
Jun	Setting up the experiment	Study designs and sample size calculations. This focus will be on what information we need for doing sample size calculations and looking at some examples. I will also introduce the approach of simulation studies at the study design phase and why they are valuable.
Sep	Systematic reviews and meta-analyses	Analysing data from your literature/ systematic review.