Demystifying statistics & programming in R

Dr Neil Burns – neil.burns@sruc.ac.uk

How the session will work

Mix of presentation are R coding

Flit between PowerPoint and R:

- Presentation and code available at my github

https://github.com/NeilMBurns

Section headings in R will be highlighted like:

```
##########################################
#####        1.  some sort of title        #####
##########################################
```

Research Interests

Population ecology and ecosystem health

- Develop methods to understand:
  - Demographic distributions of marine species
  - Spatial and temporal life stage connectivity
  - Relationship between connectivity and ecosystem health

Research Interests

PLOS ONE

RESEARCH ARTICLE

A method to improve fishing selectivity through age targeted fishing using life stage distribution modelling

Neil M. Burns[1,2]*, David M. Bailey[1], Peter J. Wright[2]

COMMUNICATIONS BIOLOGY

Otolith chemoscape analysis in whiting links fishing grounds to nursery areas
Neil M. Burns, Charlotte R. Hopkins, David M. Bailey and Peter J. Wright *2020*

https://cran.r-project.org/

The Comprehensive R Archive Network

CRAN
Mirrors
What's new?
Task Views
Search

About R

Download and Install R

Precompiled binary distributions of the base system and contributed packages, **Windows and Mac** users most likely want one of these versions of R:

- Download R for Linux
- Download R for (Mac) OS X
- Download R for Windows

R is part of many Linux distributions, you should check with your Linux package management system in addition to the link above.

Source Code for all Platforms

Windows and Mac users most likely want to download the precompiled binaries listed in the upper box, not the source code. The sources have to be compiled before you can use them.

R for Windows

CRAN
Mirrors
What's new?
Task Views
Search

About R
R Homepage
The R Journal

Subdirectories:

base	Binaries for base distribution. This is what you want to **install R for the first time**.
contrib	Binaries of contributed CRAN packages (for R >= 2.13.x; managed by Uwe Ligges). There is also information on third party software available for CRAN Windows services and corresponding environment and make variables.
old contrib	Binaries of contributed CRAN packages for outdated versions of R (for R < 2.13.x; managed by Uwe Ligges).
Rtools	Tools to build R and R packages. This is what you want to build your own packages on Windows, or to build R itself.

Please do not submit binaries to CRAN. Package developers might want to contact Uwe Ligges directly in case of questions / suggestions related to Windows binaries.

You may also want to read the R FAQ and R for Windows FAQ.

Note: CRAN does some checks on these binaries for viruses, but cannot give guarantees. Use the normal precautions with downloaded executables.

R-4.0.3 for Windows (32/64 bit)

CRAN
Mirrors
What's new?
Task Views
Search

About R
R Homepage
The R Journal

Software
R Sources

Download R 4.0.3 for Windows (85 megabytes, 32/64 bit)

Installation and other instructions
New features in this version

If you want to double-check that the package you have downloaded matches the package distributed by CRAN, you can compare the md5sum of the .exe to the fingerprint on the master server. You will need a version of md5sum for windows: both graphical and command line versions are available.

Frequently asked questions

- Does R run under my version of Windows?
- How do I update packages in my previous version of R?
- Should I run 32-bit or 64-bit R?

Please see the R FAQ for general information about R and the R Windows FAQ for Windows-specific information.

Resources

DOWNLOAD

Download RStudio Desktop

R script set up…oh yeah and making it look cool!

```
##########################################
#####        1. Setting up your script        #####
##########################################
```

START: Assemble your data as laid out in Tables 1 and 2 in the main text
Step 1: test your data for normality using ANDERSON-DARLING, SHAPIRO-WILKS or KOLMOGOROV-SMIRNOV. Are the data normally distributed?
Yes
No
Step 2: Screen the data for mis-keyed data
Step 3: attempt to normalise the data: BOX-COX procedure
Are the data now normally distributed?
Yes
No
Correct for body size
Consult a trained statistician before pursuing this analysis route
Step 4: do you want to analyse the whole animal data or correct for body size differences?
Correct for body size
Whole animal analysis
Step 5: How many treatments do you have – one or more than one?
One
More than one
Step 6: How many levels of treatment do you have – one or more than one?
More than one
One
Step 7: was each individual exposed to both treatment and control conditions?
No
Yes
Step 8: was each individual exposed to all levels?
No
Yes
Step 12: was each individual exposed to all treatments?
No
Yes
Steps 13 & 14: If 2 treatments, analyse data using TWO-WAY ANOVA. If more than 2 treatments, analyse using GLM
Steps 13 & 14: If 2 treatments then analyse data using RM TWO-WAY ANOVA. If more than 2 treatments analyse using GLM
Step 22: do you have data on the following:
a) Body masses
b) Lean and fat masses
c) Multiple organ masses
Lean and fat masses
Body masses
Multiple organ masses
Step 37: construct CORRELATION MATRIX to evaluate if organ masses are independent
Are organ masses independent of each other?
Yes
No
Step 38: perform PCA. Extract scores on main components
Step 23: Perform ANCOVA or GLM with body mass as the covariate
Step 32: Perform ANCOVA or GLM with two or more predictors. Use raw data or scores from PCA
Only 1 covariate significant
More than 1 covariate significant
Steps 24-31 INTERPRETATION OF POTENTIAL OUTCOMES (see main text for details)
Steps 33-36 INTERPRETATION OF POTENTIAL OUTCOMES (see main text for details)
Step 15: do you want to analyse the whole animal data or correct for body size differences?
Whole animal analysis
Step 16: How many treatments do you have – one or more than one?
One
More than one
Step 17: How many levels of treatment do you have – one or more than one?
One
Step 18: was each individual exposed to both treatment and control conditions?
Yes
No
Step 19: was each individual exposed to all treatment and control conditions or all levels of the treatment?
No
Yes
Analyse data using RM ANOVA
Step 9: Some treatment effects significant
Analyse data using KRUSKAL-WALLIS ANOVA
Analyse data using ONE-WAY ANOVA
Nothing significant
Some treatment effects significant
Yes
Analyse data using FRIEDMAN RM ANOVA
Step 20: Some treatment effects significant
No
Yes
Analyse data using PAIRED T-TEST
Step 10: perform post-hoc tests to locate differences: TUKEY'S TEST, DUNCAN'S TEST or PAIRED T-TESTS for RM ANOVA
Perform post-hoc tests to locate differences: TUKEY'S TEST, DUNCAN'S TEST etc., or JOHNSON-NEYMAN TEST
Analyse data using TWO-SAMPLE T-TEST
Nothing significant
Step 21: perform post-hoc tests to locate differences: MANN-WHITNEY U or WILCOXON MATCHED-PAIRS TEST
Step 11: Perform post-hoc POWER ANALYSIS
Yes
Analyse data using WILCOXON MATCHED-PAIRS TEST
No
Analyse data using MANN-WHITNEY U-TEST
Step 39: END
We strongly advise discussing your analysis with a professional statistician before publication.

START:
Assemble your data as laid out in Tables 1 and 2 in the main text
Step 1: test your data for normality using ANDERSON-DARLING, SHAPIRO-WILKS or KOLMOGOROV-SMIRNOV. Are the data normally distributed?
Yes
No
Step 2: Screen the data for mis-keyed data
Step 3: attempt to normalise the data: BOX-COX procedure
Are the data now normally distributed?
Yes
No
Correct for body size
Step 4: do you want to analyse the whole animal data or correct for body size differences?
Correct for body size
Consult a trained statistician before pursuing this analysis route
Whole animal analysis
Step 5: How many treatments do you have – one or more than one?
One
More than one
Step 6: How many levels of treatment do you have – one or more than one?
One
More than one
Step 7: was each individual exposed to both treatment and control conditions?
No
Yes
Step 8: was each individual exposed to all levels?
No
Yes
Step 12: was each individual exposed to all treatments?
No
Yes
Step 15: do you want to analyse the whole animal data or correct for body size differences?
Whole animal analysis
Step 16: How many treatments do you have – one or more than one?
One
More than one
Step 17: How many levels of treatment do you have – one or more than one?
One
Step 18: was each individual exposed to both treatment and control conditions?
Yes
No
Step 19: was each individual exposed to all treatment and control conditions or all levels of the treatment?
No
Yes
Analyse data using KRUSKAL-WALLIS ANOVA
Analyse data using FRIEDMAN RM ANOVA
Analyse data using PAIRED T-TEST
Analyse data using TWO-SAMPLE T-TEST
Analyse data using WILCOXON MATCHED-PAIRS TEST
Analyse data using MANN-WHITNEY U-TEST
Nothing significant
Step 20: Some treatment effects significant
No
Yes
Step 21: perform post-hoc tests to locate differences: MANN-WHITNEY U or WILCOXON MATCHED-PAIRS TEST
Steps 13 & 14: If 2 treatments, analyse data using TWO-WAY ANOVA. If more than 2 treatments, analyse using GLM
Analyse data using RM ANOVA
Analyse data using ONE-WAY ANOVA
Step 13 & 14: If 2 treatments then analyse data using RM TWO-WAY ANOVA. If more than 2 treatments analyse using GLM
Step 9: Some treatment effects significant
Nothing significant
Step 10: perform post-hoc tests to locate differences: TUKEY'S TEST, DUNCAN'S TEST or PAIRED T-TESTS for RM ANOVA
Step 11: Perform post-hoc POWER ANALYSIS
Step 22: do you have data on the following:
a) Body masses
b) Lean and fat masses
c) Multiple organ masses
Multiple organ masses
Lean and fat masses
Body masses
Step 23: Perform ANCOVA or GLM with body mass as the covariate
Step 32: Perform ANCOVA or GLM with two or more predictors. Use raw data or scores from PCA
Step 37: construct CORRELATION MATRIX to evaluate if organ masses are independent
Are organ masses independent of each other?
Yes
No
Step 38: perform PCA. Extract scores on main components
Only 1 covariate significant
More than 1 covariate significant
Steps 24-31
INTERPRETATION OF POTENTIAL OUTCOMES (see main text for details)
Steps 33-36
INTERPRETATION OF POTENTIAL OUTCOMES (see main text for details)
Some treatment effects significant
Perform post-hoc tests to locate differences: TUKEY'S TEST, DUNCAN'S TEST etc., or JOHNSON-NEYMAN TEST
Step 39: END
We strongly advise discussing your analysis with a professional statistician before publication.

They are just robots

Stained glass window Latin square, honouring [Ronald Fisher]

Why use someone else's robot?

Linear models (lm)
Generalised Linear Models (glm)
Generalised Additive Models (gam)
Mixed effects versions of (lmm, glmm, gamm)

Statistical modelling using glms

Why R?

But…we are getting ahead of ourselves.

	A	B	C	D
1	Site	Coral_colour	Perc_cover	S_abund
2	1A	Blue	20.5	23
3	1A	Green	9.5	2
4	1A	Red	70	19
5	1B	Blue	50	18
6	1B	Green	25.5	22
7	1B	Red	25.5	6

- Data into R from a .csv file
- Each column is a variable (explanatory (x) and response (y))
- Keep the column names short and simple but meaningful (to you)
- No spaces in column names_use_as_separator
- Be consistent with capitals

The very 1st thing you need to do = UNDERSTAND YOUR DATA!!!!!

Our toy example

Coral percentage cover (x_1)

0 to 100%

Abundance of sharks (y_1) ~

It's a whole number (<u>or maybe not</u>)
They are big and there are not loads of them so probably less than 60

Coral colour (x_2)

Blue, Brown & green

Shark abundance
Coral percentage cover

Shark abundance
Coral percentage cover

Back to R

```
##########################################
#####          2. Understanding our data     #####
##########################################
```

Building useful robots and making inferences

Build the model from the knowledge of your data then use information theory to select models (and make inferences).

A. Model selection
   - Picking the best of a bad bunch.

B. Model validation
   - So how rubbish is it?

By using AIC we are selecting the most “cost” effective model. With cost being measured in degrees of freedom.

AIC balance between over and underfitting by estimating out-of-sample deviance without needing to do cross-validation (the gold standard)

A - Model selection

Using AIC (log-likelihood ratio tests are also useful for those who like a p-value)

<u>Selection recipe</u>

- Start with the most complex model and work “back” towards the most simple

- Use AIC to choose (3 rules)
  1. Simple models are best
  2. Small AIC is best
  3. If these rules contradict (ie the more complex model has smaller AIC) then AIC should be different by more then 2

“I used backwards stepwise model selection to …”

Weirdly it’s not an exact science… (B)

General aim of model validation is:

to have normally distributed residuals

and

no obvious patterns in the residuals

WTF is a residual?

Final interpretation

Simulation allows us to interpret our complex models

A word about coefficients and factors >>>>>>>>

$$Y = mX + c \quad (+ er)$$

gradient intercept

$$Y = mX_1 + mX_2 + c \quad (+ er)$$

Our mod3

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.929914   0.417166   4.626 1.14e-05 ***
perc_cov    0.116220   0.007622  15.248  < 2e-16 ***
---
```

Our mod2

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.866607   0.513668  11.421  < 2e-16 ***
perc_cov      0.094528   0.008575  11.024  < 2e-16 ***
cor_colBrown -0.986790   0.143164  -6.893 5.75e-10 ***
cor_colGreen -0.580381   0.111171  -5.221 1.03e-06 ***
---
```

Our mod1

```
Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)           5.49784    0.60079   9.151 1.17e-14 ***
perc_cov              0.10071    0.01005  10.026  < 2e-16 ***
cor_colBrown         -1.61468    1.00986  -1.599  0.11320
cor_colGreen          3.69075    1.38125   2.672  0.00889 **
perc_cov:cor_colBrown 0.01524    0.02008   0.759  0.44973
perc_cov:cor_colGreen -0.08092    0.02581  -3.135  0.00229 **
---
```

Final interpretation – a word about staring into the abyss

```
Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)            5.49784    0.60079   9.151 1.17e-14 ***
perc_cov               0.10071    0.01005  10.026  < 2e-16 ***
cor_colBrown          -1.61468    1.00986  -1.599  0.11320
cor_colGreen           3.69075    1.38125   2.672  0.00889 **
perc_cov:cor_colBrown  0.01524    0.02008   0.759  0.44973
perc_cov:cor_colGreen -0.08092    0.02581  -3.135  0.00229 **
---
```

= 42

Back to R

```
##########################################
#####          3. Models and inference          #####
##########################################
```

Ahh wait, but what if…?

$$Y = mx + c \quad (+ er)$$

This time this one is not fixed

In this case we use GAM

Ahh wait, but what if…?

If we have repeated measures, then mixed effects model.
These models share information between groups but don't treat them separately
(ie as a factor in a model).

Ahh wait, but what if…?

If we have repeated measures, then mixed effects model.
These models share information between groups but don't treat them separately
(ie as a factor in a model).

Ahh wait, but what if…?

If we have repeated measures, then mixed effects model.
These models share information between groups but don't treat them separately
(ie as a factor in a model).

Summary and questions

- Graph and understand your data
- Avoid spending too long down rabbit holes
- Use the tools we have discussed to build the robot
  - and the “recipe” to perform model selection
- We can use the *PGR Stats in R team* to discuss and resolve problems