To call in the statistician after the experiment is done may be no more than asking him to perform a post-mortem examination: he may be able to say what the experiment died of.

Sir Ronald Fisher

Presidential Address to the First Indian Statistical Congress, 1938. Sankhya 4, 14-17

(BUT maybe you don't need to call in a statistician if you know some principles)

When assumptions are not met

Timothée Bonnet

Biological Data Science Institute

May 2, 2019

If you haven't enough R in your life

https://www.meetup.com/rladies-canberra/events/dvrjwqyzhbjb/

6 MAY

Monday, May 6, 2019

Tidying and plotting in R

Hosted by Petra K. and 2 others
From R-Ladies Canberra
Public group

Are you going? 18 people going

Share:

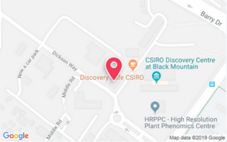
Details

R is primarily known as a tool for statistics and data analysis, but it is also great for "simpler" data wrangling and plotting purposes. As R is becoming more accessible and useful in this regards, it is becoming a more viable alternative to Microsoft Excel and other spreadsheet-based applications for wrangling,

Monday, May 6, 2019
1:00 PM to 2:30 PM
Every week on Monday until May 12, 2019

CSIRO - Synergy Building
N Science Rd · Acton

How to find us
Stringybark, Building 801, Cnr of Dickson Way and, N Science Rd, Acton ACT 2601



Take-home messages

1. Assumptions are about math properties, but also about scientific logic

Take-home messages

1. Assumptions are about math properties, but also about scientific logic
2. In general assumptions are not perfectly met

Take-home messages

1. Assumptions are about math properties, but also about scientific logic
2. In general assumptions are not perfectly met
3. There are many technics to get closer

Take-home messages

1. Assumptions are about math properties, but also about scientific logic
2. In general assumptions are not perfectly met
3. There are many technics to get closer
4. "there is no free lunch", every model makes assumptions, explicit or implicit

Take-home messages

1. Assumptions are about math properties, but also about scientific logic
2. In general assumptions are not perfectly met
3. There are many technics to get closer
4. "there is no free lunch", every model makes assumptions, explicit or implicit
5. All models are wrong, a few are useful

Statistical inference: General approach

1. Scientific question

Statistical inference: General approach

1. Scientific question

↓

2. Model and Statistical question

Statistical inference: General approach

1. Scientific question

$\downarrow$

2. Model and Statistical question $\longleftrightarrow$ 3. Data collection

Statistical inference: General approach

1. Scientific question

$\downarrow$

2. Model and Statistical question $\longleftrightarrow$ 3. Data collection

$\downarrow$

4.a Estimation

4.b Uncertainty and statistical significance

Statistical inference: General approach

1. Scientific question

2. Model and Statistical question ⟷ 3. Data collection

4.a Estimation

4.b Uncertainty and statistical significance

5. Prediction, diagnostic, check assumptions

Until "good enough"

Statistical inference: General approach

1. Scientific question

2. Model and Statistical question ⟷ 3. Data collection

4.a Estimation

4.b Uncertainty and statistical significance

5. Prediction, diagnostic, check assumptions

Until "good enough" to trust 4.a and 4.b

Statistical inference: General approach

1. Scientific question

2. Model and Statistical question ⟷ 3. Data collection

4.a Estimation

4.b Uncertainty and statistical significance

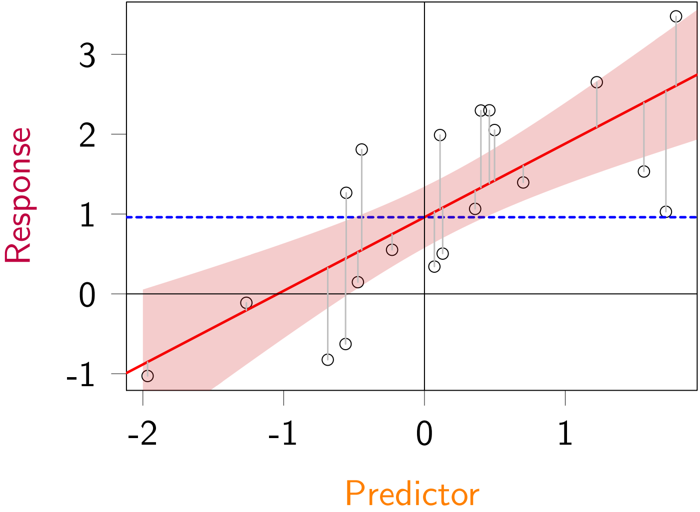
5. Prediction, diagnostic, check assumptions

Until "good enough" to trust 4.a and 4.b

6. Interpret and think about the biology

Simple linear model

Response = **Intercept** + **Slope** × **Predictor** + **Error**



What are linear model assumptions?

Linear model assumptions

Assumptions do NOT include:

- Relationship is a straight line ("Linear" means a line on some scale, not any scale)
- Data normality (Only error normality)
- Collinearity (Changes parameter meaning, not the validity)

Linear model assumptions

- Linear combination of parameters (including transformation, polynoms, interactions. . .)
  Risk: biologically meaningless
- Predictor not **perfectly** correlated
  Risk: Model won't run, unstable convergence, or huge SE
- Little error in predictors
  Risk: bias estimates (underestimate with Gaussian error)
- Gaussian **error** distribution
  Risk: Poor predictions, wrong uncertainty
- Homoscedasticity (constant error variance)
  Risk: Poor predictions, wrong uncertainty
- Independence of error
  Risk: Bias and over-optimistic uncertainty

Exercise 1

Residual distributions

Exercise 1

Exercise 2

Exercise 3

Non-independence

Exercise 3

Exercise 4

Let's think about more complex issues

Problems / Detection / Typical solutions:

- Non-Linear relationships / Thinking
  Solution: do you really need that? / non-linear models

Problems / Detection / Typical solutions:

- Non-Linear relationships / Thinking
  Solution: do you really need that? / non-linear models
- Predictor **perfectly** correlated / Singular fit
  Solution: Drop redundant variable

Problems / Detection / Typical solutions:

- Non-Linear relationships / Thinking
  Solution: do you really need that? / non-linear models
- Predictor **perfectly** correlated / Singular fit
  Solution: Drop redundant variable
- Error in predictors / Thinking / multiple measurements
  Solution: Better experiment / measurement error model

Problems / Detection / Typical solutions:

- Non-Linear relationships / Thinking
  Solution: do you really need that? / non-linear models
- Predictor **perfectly** correlated / Singular fit
  Solution: Drop redundant variable
- Error in predictors / Thinking / multiple measurements
  Solution: Better experiment / measurement error model
- Non-Gaussian **error** distribution / Plot / experimental design
  Solution: Transformation / GLM

Problems / Detection / Typical solutions:

- Non-Linear relationships / Thinking
  Solution: do you really need that? / non-linear models
- Predictor **perfectly** correlated / Singular fit
  Solution: Drop redundant variable
- Error in predictors / Thinking / multiple measurements
  Solution: Better experiment / measurement error model
- Non-Gaussian **error** distribution / Plot / experimental design
  Solution: Transformation / GLM
- Changing error variance Plot / experimental design
  Solution: Transformation / GLM

Problems / Detection / Typical solutions:

- Non-Linear relationships / Thinking
  Solution: do you really need that? / non-linear models
- Predictor **perfectly** correlated / Singular fit
  Solution: Drop redundant variable
- Error in predictors / Thinking / multiple measurements
  Solution: Better experiment / measurement error model
- Non-Gaussian **error** distribution / Plot / experimental design
  Solution: Transformation / GLM
- Changing error variance Plot / experimental design
  Solution: Transformation / GLM
- Correlated errors Thinking (/plot)
  Solution: Experimental design / control variables / mixed models

Take-home messages

1. Assumptions are about math properties, but also about scientific logic
2. In general assumptions are not perfectly met
3. There are many technics to get closer
4. "there is no free lunch", every model makes assumptions, explicit or implicit
5. All models are wrong, a few are useful