Statistics in R

Alex Douglas

a.douglas@abdn.ac.uk

@Scedacity

UNIVERSITY OF ABERDEEN

1

goals

- introduce you to some basic statistics in R
- focus on linear models that you've met previously
- fitting simple linear models in R
- linear model validation techniques in R

2

statistics in R

- many, many statistical tests available in R
- range from the simple to the highly complex
- many are included in standard installation
- you can extend the range of statistics by installing packages

3

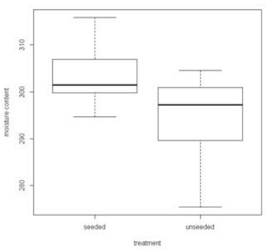
statistics in R

- example
  - does seeding clouds with dimethylsulphate alter the moisture content of clouds (can we make it rain!)
  - 10 random clouds were seeded and 10 random clouds unseeded
  - what's the null hypothesis?
  - no difference in mean moisture content between seeded and unseeded clouds

4

statistics in R

- example
  - plot these data
  - interpretation?
  - what type of statistical test do you want to use?




```
str(clouds)
'data.frame':  20 obs. of  2 variables:
 $ moisture : num  301 302 299 316 307 ...
 $ treatment: Factor w/ 2 levels "seeded","unseeded":...
```

5

statistics in R

- two sample t-test to compare the means of seeded group and unseeded group

```
t.test(clouds$moisture~clouds$treatment, var.equal=TRUE)


        Two Sample t-test

data:  moisture by treatment
t = 2.5404, df = 18, p-value = 0.02051
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  1.482679 15.657321
sample estimates:
  mean in group seeded mean in group unseeded
                303.63                 295.06
```

reject or fail to reject the null Hypothesis?

6

statistics in R

- biological interpretation?
- assumptions?
  - normality within each group
  - equal variance between groups
- test for normality with Shapiro-wilk test for each group separately

```
shapiro.test(clouds$moisture[clouds$treatment=="seeded"])
shapiro.test(clouds$moisture[clouds$treatment=="unseeded"])
```

7

statistics in R

- null hypotheses?

```
Shapiro-Wilk normality test

data:  moisture[treatment == "seeded"]
W = 0.9392, p-value = 0.544

shapiro.test(moisture[treatment=="unseeded"])

        Shapiro-Wilk normality test

data:  moisture[treatment == "unseeded"]
W = 0.8716, p-value = 0.1044
```

- I will give you a much better way of assessing normality later

8

statistics in R

- equal variance using an F test
- null hypothesis?

```
var.test(clouds$moisture~clouds$treatment)

        F test to compare two variances

data:  moisture by treatment
F = 0.5792, num df = 9, denom df = 9, p-value = 0.4283
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.1438623 2.3318107
sample estimates:
ratio of variances
         0.5791888
```

- I will give you a much better way of assessing equal variance later

9

linear models in R

- an alternative, but equivalent approach is to use a linear model to compare the means in each group
- general linear models are generally thought of as simple models, but can be used to model a wide variety of data and exp. designs
- traditionally statistics is performed (and taught) like using a recipe book (ANOVA, t-test, ANCOVA etc)
- general linear models provide a coherent and theoretically satisfying framework on which to conduct your analyses

10

what are linear models?

- thankfully, many of the statistical test you have learned so far are examples of linear models
  - one sample *t*-test
  - two sample *t*-test
  - paired *t*-test
  - ANOVA
  - ANCOVA
  - correlation
  - linear regression
  - multiple regression
  - *F*-tests
  - etc…

all cases of the (univariate) General Linear Model

11

model formulae

- general linear modelling is based around the concept of model formulae

  response variable ~ explanatory variable(s) + error

- literally read as ‘variation in response variable modelled as a function of the explanatory variable(s) plus variation not explained by the explanatory variables’
- it is the attributes of the response and explanatory variables that determines the type of linear model fitted

y ~ x	if y and x are continuous then simple linear regression
y ~ sex	If x is a categorical (nominal) variable then one-way ANOVA

12

linear modelling in R

- the function for carrying out linear regression in R is `lm`
- the response variable comes first, then the tilde ~ then the name of the explanatory variable

```
clouds.lm <- lm(moisture ~ treatment, data=clouds)
```

- how does R know that you want to perform a t-test (ANOVA)?

```
  class(clouds$treatment)
[1] "factor"
```

- here the explanatory variable is a factor.

13

linear modelling in R

- to display the ANOVA table

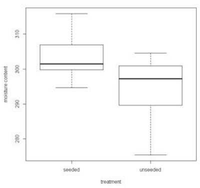
```
anova(clouds.lm)

Analysis of Variance Table

Response: moisture
          Df  Sum Sq Mean Sq F value  Pr(>F)
treatment  1  367.22  367.22  6.4538 0.02051 *
Residuals 18 1024.20   56.90
---
```

- do you notice anything about the *p* value?

14

linear modelling in R

- we have sufficient evidence to reject the null hypothesis (as before)
- therefore, there is a significant difference in the mean moisture content between clouds that were seeded and unseeded clouds
- do we accept this inference?
- what about assumptions?
- we could use Shapiro-wilks and *F* tests as before
- much better to asses visually by plotting the residuals



15

linear modelling in R

- because `clouds.lm` is a model object we can do stuff with it
- we can use the `plot()` function directly

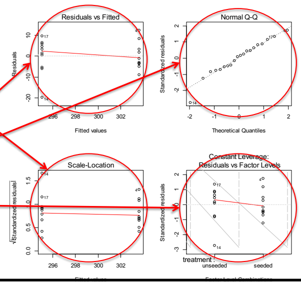
```
par(mfrow=c(2,2))
plot(clouds.lm)
```

assess equal variance assumption

assess normality assumption

assess any unusual observations

- look ok?



16

linear modelling in R

- the two sample t-test and a linear model with a categorical explanatory variable with 2 levels are equivalent
- this concept can easily be extended

difference?

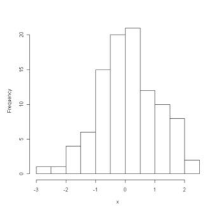
Traditional name	Model formula	R code
Bivariate regression	Y~X_1 (continuous)	lm(Y~X)
One way ANOVA	Y~X_1 (categorical)	lm(Y~X)
Two-way ANOVA	Y~X_1 (cat) +X_2(cat)	lm(Y~X_1 + X_2)
ANCOVA	Y~X_1 (cat) +X_2(cont)	lm(Y~X_1 + X_2)
Multiple regression	Y~X_1 (cont) +X_2(cont)	lm(Y~X_1 + X_2)
Factorial ANOVA	Y~X_1 (cat) * X_2(cat)	lm(Y~X_1 * X_2) or lm(Y~X_1 + X_2 + X_1:X_2)

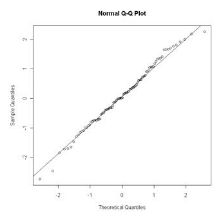
equivalent

17

extra detail of model validation in R

- normality of residuals
  - use histograms and Q-Q plots of the residuals



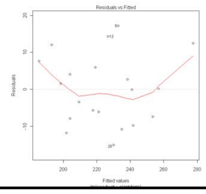


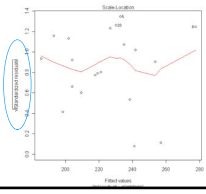
- normality is important, but linear models are fairly robust from smallish departures from normality

18

model validation in R

- equal variance (homogeneity, homoscedasticity)
  - plots of residuals (left) against fitted values and scale-location plots (right)
  - what you want to see is a complete scatter of points (sky at night)
  - what you don't want to see is any pattern or structure

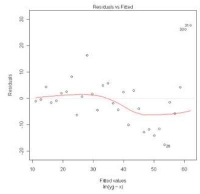


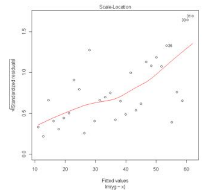


19

model validation in R

- a common problem is that the variance increases with the mean
- results in an expanding fan shaped pattern of residuals (left)
- an increase in scatter and value of standardised residuals (right)

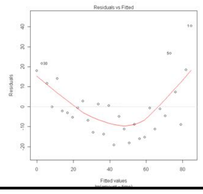


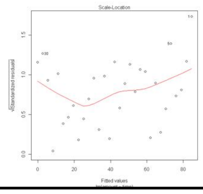


20

model validation in R

- can also detect non linearity between response and explanatory variable(s) that has not been accounted for with the structure of the model
- residuals versus fitted clearer than scale-location plot





21