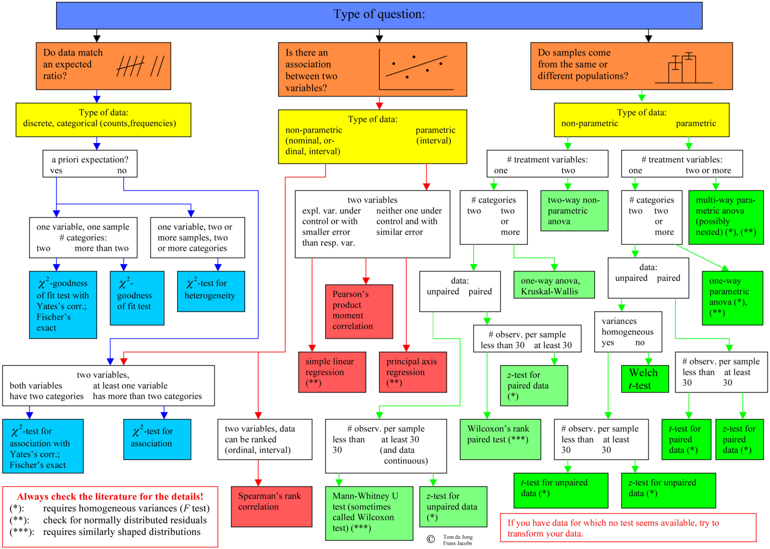
Introduction to statistical modelling with R

Conor Goold

Faculty of Biosciences,
Norwegian University of Life Sciences

02/11/2017

What we'll avoid



Type of question:

Do data match an expected ratio?

Type of data: discrete, categorical (counts,frequencies)

a priori expectation? yes no

one variable, one sample # categories: two more than two

χ^2-goodness of fit test with Yates's corr.; Fischer's exact

χ^2-goodness of fit test

one variable, two or more samples, two or more categories

χ^2-test for heterogeneity

two variables, both variables have two categories; at least one variable has more than two categories

χ^2-test for association with Yates's corr.; Fischer's exact

χ^2-test for association

Is there an association between two variables?

Type of data: non-parametric (nominal, ordinal, interval) parametric (interval)

two variables: expl. var. under control or with smaller error than resp. var.; neither one under control and with similar error

Pearson's product moment correlation

simple linear regression (**)

principal axis regression (**)

two variables, data can be ranked (ordinal, interval)

Spearman's rank correlation

Do samples come from the same or different populations?

Type of data: non-parametric parametric

treatment variables: one two

categories two two or more

two-way non-parametric anova

data: unpaired paired

one-way anova, Kruskal-Wallis

observ. per sample less than 30 at least 30

z-test for paired data (*)

Wilcoxon's rank paired test (***)

observ. per sample less than 30 at least 30 (and data continuous)

Mann-Whitney U test (sometimes called Wilcoxon test) (***)

z-test for unpaired data (*)

treatment variables: one two or more

categories two two or more

multi-way parametric anova (possibly nested) (*), (**)

data: unpaired paired

one-way parametric anova (*), (**)

variances homogeneous yes no

Welch *t*-test

observ. per sample less than 30 at least 30

t-test for paired data (*)

z-test for paired data (*)

observ. per sample less than 30 at least 30

t-test for unpaired data (*)

z-test for unpaired data (*)

Always check the literature for the details!

(*): requires homogeneous variances (*F* test)

(**): check for normally distributed residuals

(***): requires similarly shaped distributions

If you have data for which no test seems available, try to transform your data.

© Tom de Jong Frans Jacobs

What we'll do instead

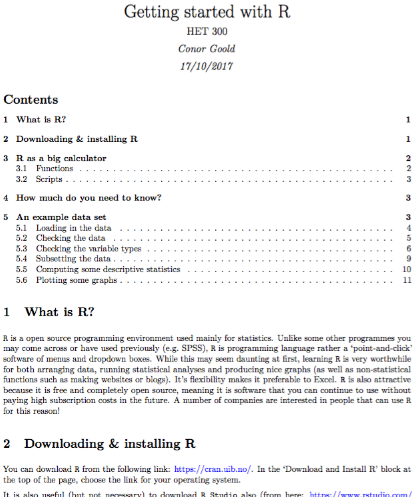
Most statistical models have the same mathematical form, which is called the **linear model (LM)**.

LMs predict one variable from a linear or additive combination of other variables.

Examples

Height = weight
Test score = age + sex + social class + ethnicity + ...

Using R



Getting started with R

HET 300

Conor Goold

17/10/2017

Contents

1 What is R?

R is a open source programming environment used mainly for statistics. Unlike some other programmes you may come across or have used previously (e.g. SPSS), R is programming language rather a 'point-and-click' software of menus and dropdown boxes. While this may seem daunting at first, learning R is very worthwhile for both arranging data, running statistical analyses and producing nice graphs (as well as non-statistical functions such as making websites or blogs). It's flexibility makes it preferable to Excel. R is also attractive because it is free and completely open source, meaning it is software that you can continue to use without paying high subscription costs in the future. A number of companies are interested in people that can use R for this reason!

2 Downloading & installing R

You can download R from the following link: https://cran.uib.no/. In the 'Download and Install R' block at the top of the page, choose the link for your operating system.

It is also useful (but not necessary) to download R Studio also (from here: https://www.rstudio.com/

Outline

Choosing your variables

Identifying the type of variable

The linear model

Presenting the results

Generalised linear model

Multiple explanatory variables

Multi-level models

Outline

Choosing your variables

Identifying the type of variable

The linear model

Presenting the results

Generalised linear model

Multiple explanatory variables

Multi-level models

The goal of statistical modelling

Is variation in one variable influenced by variation in other variables?

Caution

Most statistical models are not about finding causal relationships

Response vs. explanatory variables

The variable we want to predict variation in is the **response** variable (also known as *dependent* or *predicted* variable).

The variable that we believe explains variation is the **explanatory** variable (also known as *independent* or *predictor* variable).

Examples

- Test scores and age

Examples

- Test scores and age
- Number of dog barks and time left alone

Examples

- Test scores and age
- Number of dog barks and time left alone
- Caffeine consumption and reaction time

Examples

- Test scores and age
- Number of dog barks and time left alone
- Caffeine consumption and reaction time
- Seconds taken for a dog to recall and training level

Examples

- Test scores and age
- Number of dog barks and time left alone
- Caffeine consumption and reaction time
- Seconds taken for a dog to recall and training level

Which variable is the response or explanatory is determined by context/research question.

You may have multiple explanatory variables

- Test scores = child age, sex and ethnicity.
- Number of dog barks = time left alone and level of separation anxiety

Important

When we have multiple explanatory variables, we are interested in the **unique** effects of each variable.

Outline

Choosing your variables

Identifying the type of variable

The linear model

Presenting the results

Generalised linear model

Multiple explanatory variables

Multi-level models

Metric, count, ordinal, nominal

- Duration (seconds) of a bark (metric)

Metric, count, ordinal, nominal

- Duration (seconds) of a bark (metric)
- Number of barks in 10 minutes (count)

Metric, count, ordinal, nominal

- Duration (seconds) of a bark (metric)
- Number of barks in 10 minutes (count)
- First, second and third loudest barks (ordinal)

Metric, count, ordinal, nominal

- Duration (seconds) of a bark (metric)
- Number of barks in 10 minutes (count)
- First, second and third loudest barks (ordinal)
- Frustrated bark, playful bark, aggressive bark (nominal)

Metric variables

Metric variables are ‘continuous’ scales, which can have decimal points, i.e. they make sense on the real number line.

Examples

- Temperature
- Response time
- Height, weight

Count variables

Count variables are a type of metric scale, but are postive integers and reflect the number of events in a certain interval (can be converted to proportions).

Examples

- Number of animals in an area
- Number of times a behaviour occurs within x minutes
- Number of people voting for different political parties

Ordinal variables

Ordinal variables reflect order, although distances between values are not always equal.

Examples

- Podium places in a race
- Likert scale responses (”Rate on a scale of 1 to 5 how much you agree with...”)
- Order of students by test scores (first, second, third...)

Nominal variables

Nominal variables reflect different categories with no natural order.

Examples

- Different political parties
- Different colour cars
- Different emotional states

When using linear models...

We usually apply a linear model appropriate for our response variable type (metric, count, ordinal, nominal).

If our explanatory variables are metric, count or ordinal, they can usually be treated as the same type of variable in statistical models. If they are nominal, they are treated different.

Summary so far

- Identify response and explanatory variables
- Identify data types
  - metric, count, ordinal or nominal
- The same general statistical model can be applied to data of all these types!

Outline

Choosing your variables

Identifying the type of variable

The linear model

Presenting the results

Generalised linear model

Multiple explanatory variables

Multi-level models

Linear equation

$$y = \alpha + \beta x$$

- y is the response variable
- α is where the line crosses the y-axis
- x is the explanatory variable
- β is the slope coefficient for x

Interpreting linear equations

When $x = 0$,

$$y = \alpha + \beta \cdot 0,$$
$$y = \alpha.$$

α is the y-intercept.

Interpreting linear equations

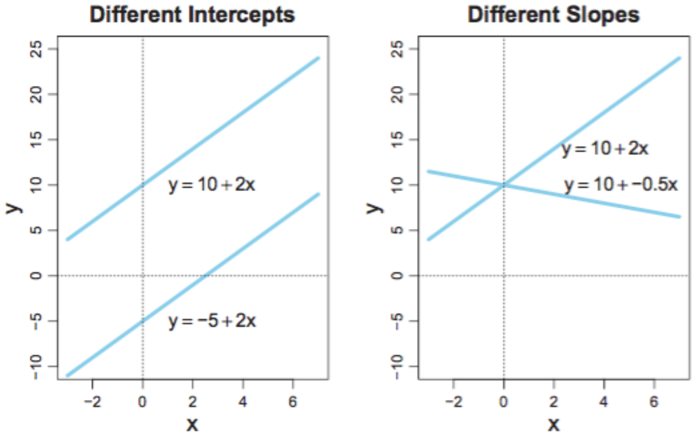
β is amount of change in y with **one unit increase** in x.

Example

If $\alpha = 3$, $\beta = 2$, and $x = 2$

$$y = \alpha + \beta x = 3 + 2 \cdot 2 = 7$$

Linear equations



From Kruschke (2014), *Doing Bayesian Data Analysis*

What are statistical models?

Statistical models are mathematical machines that produce data.
Our goal is to ‘fit’ a model that matches our data.

Why do we need statistical models?

Parameters

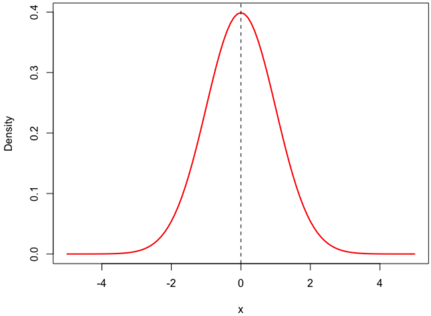
The values for α and β (parameters) are **unknown**

1. We collect a sample of data from the population
2. It's the statistical model's job to find out the most plausible values of those parameters

Linear regression with normal distributions

Linear models predict the ‘central tendency’ (e.g. mean) of the data, and how the central tendency changes when predictor variables change.

The simplest linear regression uses a **normal distribution** to describe how the data is distributed around the mean.



Linear regression with normal distributions

$$y \sim Normal(\mu, \sigma)$$

$$\mu = \alpha + \beta x$$

In English

y is normally distributed with mean μ and standard deviation σ.

The mean μ is a linear function of an intercept and the influence of x.

y **is normally distributed around its mean.**

Example: hours hunting and age

What's the relationship between hours spent hunting and age?

$$hours \sim Normal(\mu, \sigma)$$

$$\mu = \alpha + \beta age$$

```
lm( hours ~ 1 + age )
```

Linear regression assumptions

- Response variable is metric (i.e. decimals make sense)

Linear regression assumptions

- Response variable is metric (i.e. decimals make sense)
- Response variable can be both positive and negative

Linear regression assumptions

- Response variable is metric (i.e. decimals make sense)
- Response variable can be both positive and negative
- Independence of data points

Linear regression assumptions

- Response variable is metric (i.e. decimals make sense)
- Response variable can be both positive and negative
- Independence of data points
- Normally distributed **residuals** (difference between the model predictions and the raw data values)

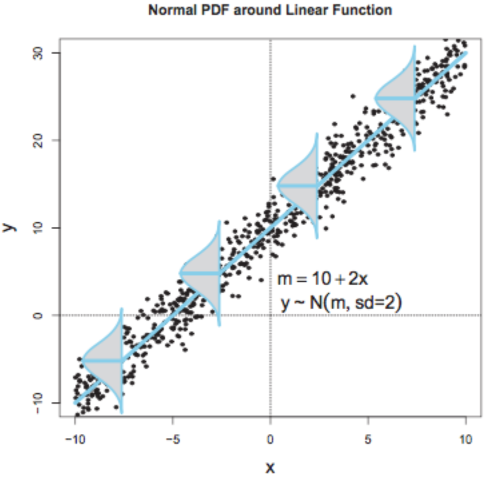
Linear regression assumptions

- Response variable is metric (i.e. decimals make sense)
- Response variable can be both positive and negative
- Independence of data points
- Normally distributed **residuals** (difference between the model predictions and the raw data values)
- Residuals should not depend on the fitted values

Linear regression assumptions

- Response variable is metric (i.e. decimals make sense)
- Response variable can be both positive and negative
- Independence of data points
- Normally distributed **residuals** (difference between the model predictions and the raw data values)
- Residuals should not depend on the fitted values
- Homogeneity of variance in residuals

Linear regression assumptions



From Kruschke (2014), *Doing Bayesian Data Analysis*

Practical

Fitting a linear regression in R

Using a nominal explanatory variable

Nominal explanatory variables can also be included in the model by using a system of dummy variables.

Example

Imagine a binary explanatory variable x, which we code as 0 (for one group) and 1 (for the other group).

When $x = 0$,

$$y = \alpha + \beta \cdot 0 = \alpha$$

When $x = 1$,

$$y = \alpha + \beta \cdot 1$$

Thus, β is now the estimated difference between the groups.

Nominal explanatory variable with $>$ than 2 levels

Example

Our nominal explanatory var. x has three levels (three groups). We create two new variables:

$x_{group2} = 1$ when an observation is in group 2, 0 otherwise.

$x_{group3} = 1$ when an observation is in group 3, 0 otherwise.

These 'dummy variables' always equal the number of levels - 1. The linear model can now be written with two β terms:

$$y = \alpha + \beta_{group2} \cdot x_{group2} + \beta_{group3} \cdot x_{group3}$$

Nominal explanatory variable with $>$ than 2 levels

When $x_{group2} = 0$ and $x_{group3} = 0$,

$$y = \alpha + \beta_{group2} \cdot 0 + \beta_{group3} \cdot 0 = \alpha$$

When $x_{group2} = 1$ and $x_{group3} = 0$,

$$y = \alpha + \beta_{group2} \cdot 1 + \beta_{group3} \cdot 0 = \alpha + \beta_{group2}$$

When $x_{group2} = 0$ and $x_{group3} = 1$,

$$y = \alpha + \beta_{group2} \cdot 0 + \beta_{group3} \cdot 1 = \alpha + \beta_{group3}$$

Practical

Fitting a linear regression with nominal explanatory variables in R

Outline

Choosing your variables

Identifying the type of variable

The linear model

Presenting the results

Generalised linear model

Multiple explanatory variables

Multi-level models

Interpreting results

First, we interpret the parameter estimates for our first linear model.

Hours model

$\alpha = 6.82$; $\beta = 0.0004$

“The number of hours spent hunting when age is 0 is 6.82 and a one year increase in age is associated with a 0.004 increase in the number of hours spent hunting.”

Interpreting results

Standard errors tell us about the uncertainty in the parameter estimates, and we can use them to find a 95% confidence interval.

Hours model

α standard error $= 0.13$,
95% CI $= 6.82 \pm 1.96 \cdot 0.13 = [6.57, 7.07]$

β standard error $= 0.003$,
95% CI $= 0.0004 \pm 1.96 \cdot 0.003 = [-0.005, 0.006]$

Determing the ‘significance’ of our results

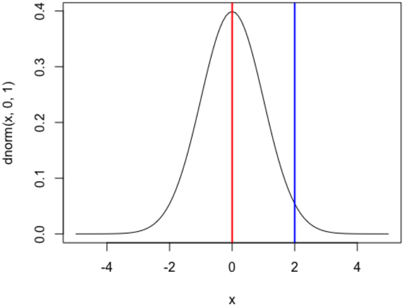
Scientists often report the statistical significance of their explanatory variable estimates. We could do this using 95% CIs (best) or *p*-values (usually need to be less than 0.05), which are reported in the last column of the model summary.

Hours model

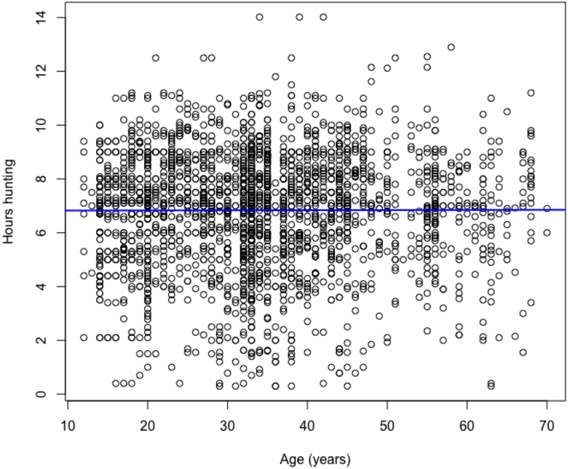
“The effect of age on hours spent hunting was not statistically significant ($\beta = 0.0004$; SE = 0.003; $p = 0.892$; 95% CI: [-0.005, 0.006])”.

What do p values really mean?

P-values represent the probability of our parameter values (e.g. β) coming from a distribution of parameter values we would expect to see if the null hypothesis was true (i.e. no effect of our explanatory variable).



Plotting the results



Practical

Presenting the results in R.

Full write up

"The influence of age on hours spent hunting was analysed using a linear regression model, with hours spent hunting as the response variable and age as the explanatory variable. A normal distribution was appropriate for the residuals based on regression diagnoistics (e.g. QQ-plots, residuals vs. fitted plots). The effect of age on hours spent hunting was not statistically significant ($\beta = 0.0004$; SE = 0.003; $p = 0.892$; 95% CI: [-0.005, 0.006])."

Outline

Choosing your variables

Identifying the type of variable

The linear model

Presenting the results

Generalised linear model

Multiple explanatory variables

Multi-level models

Non-normal residuals or non-metric data

The assumptions of the linear model may not be adequate if:

- Your response variable is not metric (e.g. is count, ordinal, or nominal)
- The residuals of your model are not normally distributed
- Your response variable is bounded (e.g. cannot be lower than 0)

Generalised linear models

In these cases, we can apply the **generalised linear model (GLM)**. It is ‘generalised’ because it generalises the linear model to different types of response variables.

GLMs convert your response variable onto a linear scale, and then fit a normal linear model as we did before. To interpret the results on the original scales, you must convert the paramter values back to the original scale.

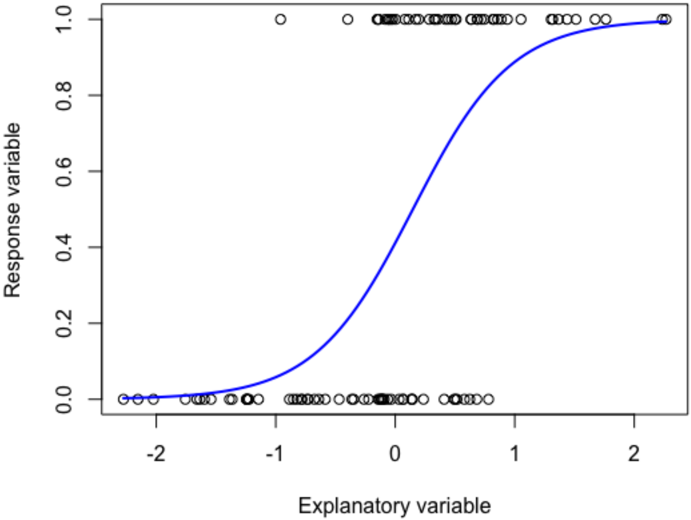
Logistic regression

Example

Imagine scan sampling a group of animals, and recording 1 if a particular behaviour is seen or 0 if not. We are interested in how the behaviour changes with time of observation.

The response variable is a series of 0s and 1s. It no longer makes sense to ask "How much does the response variable change with each increase in the explanatory variable". But we can ask "**How does the probability of the response variable change with each increase in the explanatory variable?**"

Logistic regression



Logistic regression

$$y \sim Bernoulli(p)$$

$$logit(p) = \alpha + \beta x$$

$$\text{where } logit(\mu) = \frac{\mu}{1-\mu}$$

In English

y is distributed as a Bernoulli variable (0s and 1s), with probability p. p is mapped to the linear scale using the logit **link function**, and is predicted by the linear equation $\alpha + \beta x$. To interpret α and β on the probability scale, the inverse logit function or logistic function needs to be applied.

Logistic regression in R

Fitting a logistic model in R.

Link functions

In GLMs, the underlying linear model is the same but the residual distribution and link function are just different. You will always find a linear model suitable for your type of response variable!

Other common types of GLM:

- Binary/binomial variable: binomial regression and logit/probit link function
- Count response variable: Poisson regression and log link function
- Right-skewed metric variable: gamma regression with log link function
- Nominal variable > 2 categories: multinomial regression with logit/probit link function

Outline

Choosing your variables

Identifying the type of variable

The linear model

Presenting the results

Generalised linear model

Multiple explanatory variables

Multi-level models

Multiple explanatory variables

We can just keep adding terms to a linear model, e.g.

$$y \sim \mathit{Normal}(\mu, \sigma)$$

$$\mu = \alpha + \beta_1 x_1 + \beta_2 x_2$$

Multiple explanatory variables

We can just keep adding terms to a linear model, e.g.

$$y \sim Normal(\mu, \sigma)$$

$$\mu = \alpha + \beta_1 x_1 + \beta_2 x_2$$

The interpretation of each β parameter is the amount of change in the response variable per unit increase in the explanatory variable, holding all other explanatory variables in the model at 0.

Outline

Choosing your variables

Identifying the type of variable

The linear model

Presenting the results

Generalised linear model

Multiple explanatory variables

Multi-level models

Multi-level models

For models with repeated measures on many different groups (e.g. individuals), we may want to account for the correlation within groups by using a multi-level, mixed effect or random effects model.

The simplest multi-level model just includes a separate intercept parameter for each group.

A simple multi-level model

$$y \sim Normal(\mu, \sigma)$$

$$\mu = \alpha + \nu_j + \beta x$$

In English

In this model, there is an extra parameter ν. If we have 2 repeated measurements on 100 individuals, where $j = 1$ to 100, the ν_j parameter ensures that each individual has their own intercept parameter, which is a deviation from the group-level intercept α. This allows us to detect the variation among individuals as well as account for the lack of independence between data points coming from the same person.