Seminar 4 - Basic models in R

Ursula Torres and Hannah Franklin

University of Canterbury 29/09/14 & Lincoln University 30/09/14

Aims

To learn how to run several basic models in R:

1. Simple linear regression
2. Multiple linear regression
3. ANOVA
4. Generalised linear models - poisson and binomial distributions

1. Simple linear regression

To test for a linear relationship between a continuous response variable (*y*) and a continuous predictor variable (*x*).

Syntax: **model.output** < - lm (**response** ~ **predictor**)

Example: The data set "tannin.txt" gives the growth of caterpillars on diets containing different amount of tannin (this is from Chapter 10, Regression, Michael Crawley's textbook "The R Book").

Is there a relationship between the amount of tannin and the growth of caterpillars?

```
#########Linear regression############

#fit the regression model
model.output<-lm(tannin_data$growth~tannin_data$tannin)

#alternatively if you specify the data name, you don't need to put $....
model.output<-lm(growth~tannin,data=tannin_data)

#what does the model.output object contain?
names(model.output)

#add a regression line to the plot
abline(model.output)

#interpret the results:
summary(model.output)
```

2. Multiple linear regression

Multiple linear regression models the relationship between two or more continuous explanatory variables (x1, x2, etc.) and a continuous response variable (y) by fitting a linear equation to observed data.

Syntax:

- Joining two predictors with a + sign fits both predictors, but not the interactions between them.
  model.output < - lm (**response** ~ **predictor1** + **predictor2**)
- If you join them with * signs you also get all the predictors and all the possible interactions (saturated model).
  model.output < - lm (**response** ~ **predictor1** * **predictor2**)
- If you want to specify particular interactions, they require a colon (:). (same as previous)
  model.output < - lm (**response** ~ **predictor1** + **predictor2** + **predictor1**:**predictor2)**

Example: The data set "ozone.data.txt" consists of daily ozone measurements and temperature, wind and solar radiation recorded (this is part of the built-in airquality data set in R).

Is there a relationship between temperature, wind, solar radiation and the interaction of these variables and ozone levels?

```
 96 #########Multiple linear regression##########
 97
 98
 99 #saturated model (all predictors+two way interactions+three way interaction)
100 model.output2<-lm(ozone~temp*wind*rad,data=pollution_ozone)
101
102 summary(model.output2) # coefficients table
103 anova(model.output2)  # main effects - check significance
104
105
106 #model with all predictors+two way interactions only
107 model.output3<-lm(ozone~temp+wind+rad+temp:wind+wind:rad+temp:rad,data=pollution_ozone)
108
109 summary(model.output3)
110 anova(model.output3) # three-way interaction is no longer there
111
112
113 #equivalent command for having model with all predictors+two way interactions only
114 model.output3<-lm(ozone~(temp+wind+rad)^2,data=pollution_ozone)
115 summary(model.output3)
```

3. ANOVA (Analysis of variance)

ANOVA analyses the effect of a categorical predictor variable (x) on a continuous response variable (y) by comparing two or more (k) group means.

Syntax: **anova.output** < - aov (**response** ~ **predictor**)

Example: The data frame "paint_data" (run R code to load this) is from a factorial experiment (two factors are crossed at every level). This investigated the effect of paint application method (dipping or spraying) and primer type (1-3) on paint adhesion force (data from Applied Statistics and Probability by Montgomery & Runger).

Does primer type, application method and the interaction (between these variables) affect paint adhesion force?

```
185 #Single factor ANOVA model
186
187 anova1<-aov(adhf ~ applic , data=paint_data)
188
189 anova(anova1) #ANOVA table
190 summary.lm(anova1) # gives a coefficients table as for as lm and R2 value
```

4. Generalised linear models

The generalised linear model is an extension of linear multiple regression and can be used when errors of the response variable are not normally distributed. GLM can also deal with continuous and categorical variables.

Syntax:

- Same as linear models but you specify which error structure you want to use.

model.output < - glm (**response** ~ **predictor1** + **predictor2,** family = **Error structure**)

Example: Poisson distribution

The data set warpbreaks is built in in R and gives the results of an experiment to determine the effect of wool type (A or B) and tension (low, medium or high) on the number of breaks per loom.

Does the tension, wool type and their interactions affect the number of breaks?

Since the response variable (breaks) is a count, the most adapted error structure is Poisson.

```
247 ###GLM poisson error
248 data(warpbreaks)
249 warpbreaks
250
251 model_breaks<-glm(breaks~wool*tension, warpbreaks, family=poisson)
252 summary(model_breaks)
```

Exercise: Binomial distribution

The data frame “distribution_trout.txt” has the distribution (presence/absence) of trout in 30 sites (data extracted from doubs data set which is built in the package ade4. The data was modified for the course purposes).

Associated to the distribution, there are environmental (predictor) variables:
dfs = distance from source
alt = altitude
slo = slope
flo = flow
pH
har = hardness of water
pho = phosphate
nit = nitrate
amm = ammonium
oxy = oxygen
anthro= level of human impact (Low, Medium and High).

Is the presence/absence of the trout related to flow, ammonium and pH and their interactions?