Likelihood ratio tests for model selection

Kieran Campbell

kieran.campbell@sjc.ox.ac.uk

January 11, 2016

1 Introduction

One way to compare two statistical models is by using the Likelihood Ratio Test (LRT). If we fit two models M_1 and M_2 to some data, how do we decide whether to favour M_1 over M_2? We will have two values for the (log-) likelihood at the maximum likelihood estimates l_1 and l_2, so in theory we could pick the model that gives the larger likelihood. However, if we increase the number of parameters in model 2 over model 1, we can *always* increase the likelihood (akin to over-fitting) and the likelihoods are themselves random variables, implying there will be statistical uncertainty as to whether $l_2 > l_1$.

The LRT provides a solution to this. It models the difference in log-likelihoods as a random variable and takes into account the number of parameters in each model. Consequently, we can say the probability of seeing a difference in log-likelihoods as extreme as we do if the models are just as good as each other.

Specifically

$$D = -2\ln(\text{likelihood for null model}) + 2\ln(\text{likelihood for alternative model}) \tag{1}$$

follows a χ^2 distribution with $\text{df}_a - \text{df}_n$ degrees of freedom (where df_a and df_n are the degrees of freedom for the alternative and null models respectively).

The rest of this tutorial follows in `R`. It is recommended you work through all the code to set yourself up for the assignment.

2 Basic example: linear regression

One application is to decide whether some data follows a linear trend or a flat one. This is equivalent to choosing between the linear models

$$y \sim \mathcal{N}(\beta_0, \sigma^2)$$

and

$$y \sim \mathcal{N}(\beta_0 + \beta_1 x, \sigma^2)$$

1. If we set $\beta_1 = 0$ the second model becomes the first. This means we can use the LRT. Why?

2. We can think of the degrees of freedom as the number of free parameters. How many degrees of freedom do the two models have? What's the difference between them?

To begin this example, install the `R` package `modelselectionworkshop`:

```
install.packages("devtools")
devtools::install_github("kieranrcampbell/modelselectionworkshop")
```

Some example data called `linearExample` is included. To load data in `R`, call `data`. To view the structure of data, call `str`:

```
library(modelselectionworkshop) # load the modelselectionworkshop package
data(linearExample)
str(linearExample)

## List of 4
##  $ x          : num [1:50] 3.96 1.12 2.73 2.24 6.41 ...
##  $ y          : num [1:50] 3.6829 -1.6282 0.2401 0.0748 9.6896 ...
##  $ null_loglik: num -144
##  $ alt_loglik : num -124
```

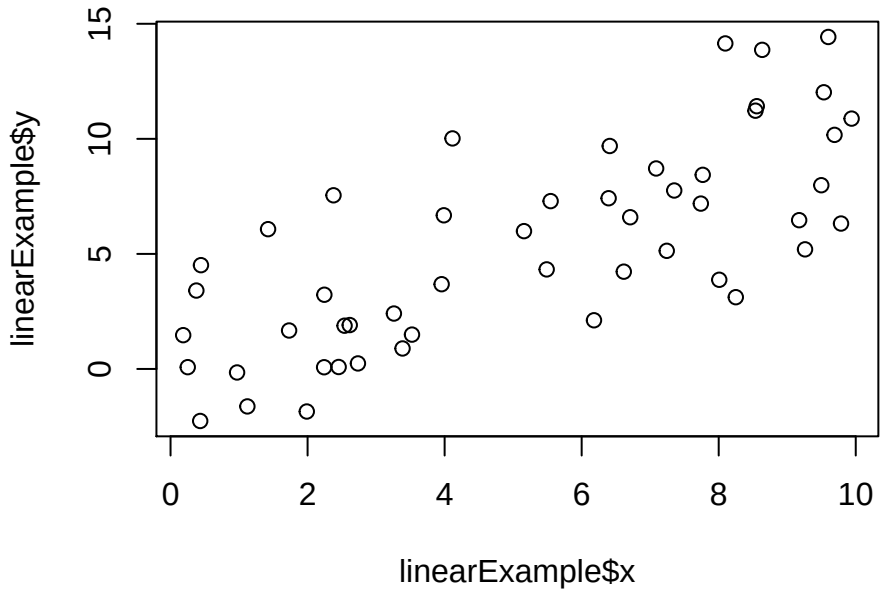
This is a `list` object in which elements can be accessed using the dollar operator. There are four elements included: `x` and `y` - samples of independent and dependent variables, and `null_loglik` and `alt_loglik` - log-likelihoods at the maximum likelihood estimates for a flat model (the null) and a linear model (the alternative).

Some useful R commands

- To get help for a command inside the `R` terminal call `?command`
- To find the dimension of a matrix/array `data` call `dim(data)`. If it's a vector then use `length(data)`
- To find the type of an object (e.g. integer or character) call `typeof(object)`. To find its class (e.g. matrix or vector) call `class(object)`

We can plot the data in `R` using the `plot` command:

```
plot(linearExample$x, linearExample$y)
```



We can form the statistic D defined in equation 1:

```
D = -2 * linearExample$null_loglik + 2 * linearExample$alt_loglik
print(D)

## [1] 39.94306
```

To find the probability of seeing a difference in log-likelihoods, we evaluate D with respect to the χ^2 distribution (in case you didn't get it earlier, the alternative model has 3 degrees of freedom, the null 2 so the χ^2 will have $3 - 2 = 1$). We do this using the `pchisq` function in `R`. Since we want the cumulative probability for all values of D *greater* than the observed, we set `lower.tail = FALSE`. Therefore, our p-value becomes:

```
pchisq(D, df = 1, lower.tail = FALSE)

## [1] 2.614749e-10
```

In standard frequentist hypothesis testing, a p-value of less than 0.05 is considered small enough to reject the null hypothesis (but, confusingly, not to *accept the alternative*), so given how small our p-value is we can be (fairly?) confident that our data follows a linear trend.

3a. (optional) What is a p-value? What *isn't* a p-value?

3b. (optional) What is the range of the likelihood function (in other words, what set of values could the likelihoods we compute possibly have)? What is the range of the log-likelihood function?

3 Biological application: pseudotime

Many experiments involve cells progressing through a biological process such as differentiation or apoptosis (cell death). Using RNA sequencing the gene expression in individual cells can be measured. As a result the expression profile of each cell represents a distinct time point through the process, known as the *pseudotime*, and algorithms have been developed to order cells and assign pseudotimes to them.

Once cells have been ordered in pseudotime it is useful to find which genes are *differentially expressed* across pseudotime (in other words, what genes change more than expected at random across pseudotime?). A simple way of doing this involves model selection: we construct a null model that relates to no differential expression and an alternative model that relates expression to some function of pseudotime, then pick the model that best explains the differential expression. If this model is the alternative model, we designate the gene as differentially expressed.

3.1 Model

Let y_{ij} denote the $\log_2$ gene expression of gene i in cell j at pseudotime t_j then

$$y_{ij}(t_j) \sim \mathcal{N}(\mu_i(t_j),\, \sigma_i^2) \tag{2}$$

where

$$\mu_i(t_j) = \begin{cases} \mu_i^{(0)}, & \text{if gene } i \text{ not differentially expressed,} \\ \frac{2\mu_i^{(0)}}{1+\exp\left(-k_i(t_j - t_i^{(0)})\right)}, & \text{if gene } i \text{ differentially expressed.} \end{cases} \tag{3}$$

4. Why is using a likelihood ratio test suitable for comparing these models?

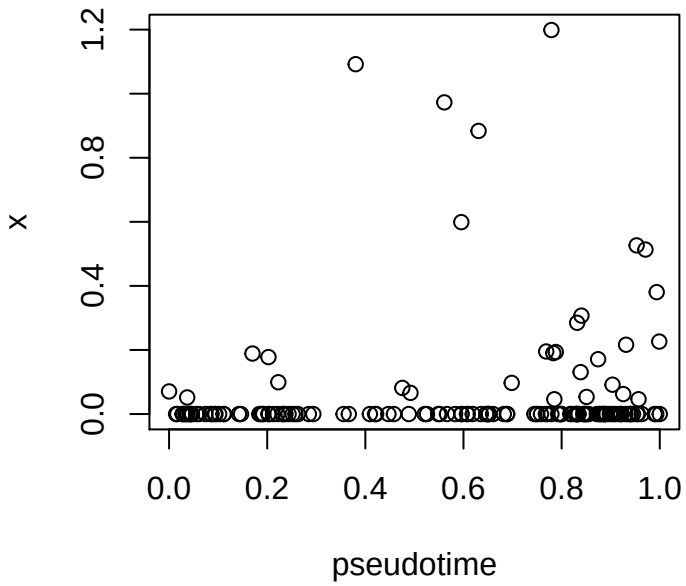
5. How many degrees of freedom does the null model have? How many does the alternative model have?

The `modelselectionworkshop` package includes a 500 by 155 matrix of gene expression measurements (for 500 genes and 155 cells) `X` and a 155 element vector `pseudotime` of pseudotimes for each cell:

```
library(modelselectionworkshop) # load package
data(X)
data(pseudotime)
str(X)

##  num [1:500, 1:155] 0 0.55 0 0.572 0.785 ...
##  - attr(*, "dimnames")=List of 2
##   ..$ : chr [1:500] "ENSG00000233750.3" "ENSG00000176595.3" "ENSG00000102098.13" "ENSG00000124222.16" .
##   ..$ : chr [1:155] "T0_CT_A01" "T0_CT_A03" "T0_CT_A05" "T0_CT_A06" ...

str(pseudotime)

##  num [1:155] 0.0742 0.1119 0.17 0.034 0.0168 ...
```

To get a feel for the data let's look at the first gene. In `R` we index matrices using square brackets, with the first value corresponding to subsetting on the row and second on the column. The `plot` command comes in handy for quickly looking at data:

```
x <- X[1,]
plot(pseudotime, x)
```



The functions `fit_null_model` and `fit_alt_model` provided by the package fit the two models and provide the log likelihoods at the maximum likelihood estimates. These return `list` objects in `R`, with two entries: `par` corresponding to the parameter vector and `loglik` corresponding to the log-likelihood. (Hint: in `R` an element named `e` in the list `L` can be accessed using the dollar sign like `L$e`.)

```
null_model <- fit_null_model(x)
alt_model <- fit_alt_model(x, pseudotime)
print(null_model)

## $par
##        mu0     sig_sq
## 0.05945317 0.03512770
##
## $loglik
## [1] 40.09383

print(alt_model)
```

```
## $par
##          mu0            k           t0       sig_sq
## 4.069274e-02 1.898407e+03 3.697681e-01 3.390793e-02
##
## $loglik
## [1] 42.33715
```

6. Find a p-value for the first gene being differentially expressed. (Hint: `pchisq` will come in handy.)

7. Find p-values for all genes in the dataset. (Hints: `apply` can be used to apply a function over a matrix. `myfunc <- function(arg1, arg2)` defines a function in `R`). If you have computed multiple p-values during an analysis, always histogram them (using `hist()`). If the null hypotheses are true for all genes the p-values should follow a uniform distribution.

8. What subsequent analysis needs done before we can decide if a gene is differentially expressed? (Hint: `p.adjust` will come in handy.)

9. What are the other weaknesses of this method?