

Trashball and logistic models

For this exercise you will need Microsoft Excel with the **Solver Add-In** installed. You may need to install the add-in, for which you may require your original MS Office installation discs.

(Unfortunately, freeware spreadsheet programs do not have the functionality of Excel's Solver.)

In Excel 2003, go to the 'Tools' menu and look for 'Solver...'. If it is not there, click on 'Add-ins...' and make sure that the "Solver Add-in" item is checked.

In Excel 2007, go to the Data tab and look for Solver in the 'Analysis' group. If it is not there:

1. Click on the Office Button.
2. Click on the 'Excel Options' button at the bottom of the window.
3. Choose 'Add-ins' from the menu on the left, select 'Excel Add-ins' in the drop-down box at the bottom of the window and click 'Go'.
4. Check the "Solver Add-in" item and click OK.

You will also need R statistical software installed.

Objective

In wildlife biology, we are often dealing with binary variables – survived/dead, detected/not detected, occupied/not occupied, etc. – and we estimate the probability of “success”, ie. survival, detection, occupancy. Often we want to investigate a relationship between the binary variable (the ‘response’) and other variables known as explanatory variables, predictors, or just covariates. The most common way to do this uses logistic modelling.

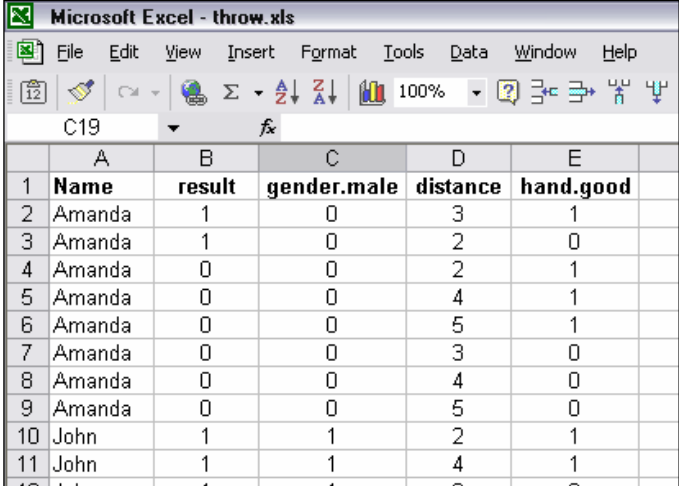
In this exercise, you try throwing the sock into a box from different distances and with either hand. We want to know if the probability of success depends on distance or which hand you use. We'll also record the gender of the thrower, as that may also have an effect. We'll analyse the data in Excel using Solver, and in R with the `glm` function.

Trashball

The name for this activity comes from the published description, Morrell & Auer (2007), where they threw balls into a trash can. We'll adapt this: maybe we should call it “Sockbox” instead.

You will try throwing the sock into the box placed against the wall from distances of 2, 3, 4 and 5m, both with your “good” hand – the one you write with – and your “bad” hand. For each throw, record whether you succeeded or not (you're allowed to bounce the sock off the wall). If the group is too large, there may not be time for each person to try all 8 combinations, so (eg) some try from 2 and 4m, others from 3 and 5m.

Record the results in a spreadsheet as shown on the right. Result = 1 means the sock went in; gender.male = 1 for males; hand.good = 1 if you were using your good hand.



The screenshot shows a Microsoft Excel spreadsheet titled 'Microsoft Excel - throw.xls'. The spreadsheet contains data for a trashball experiment. The columns are labeled 'Name', 'result', 'gender.male', 'distance', and 'hand.good'. The rows show data for Amanda and John.

	A	B	C	D	E
1	Name	result	gender.male	distance	hand.good
2	Amanda	1	0	3	1
3	Amanda	1	0	2	0
4	Amanda	0	0	2	1
5	Amanda	0	0	4	1
6	Amanda	0	0	5	1
7	Amanda	0	0	3	0
8	Amanda	0	0	4	0
9	Amanda	0	0	5	0
10	John	1	1	2	1
11	John	1	1	4	1
12	John	1	1	2	0

Logistic models

The simplest model connecting a response variable (y) to covariates (x_1, x_2, \dots) is a **linear model**:

$$y = \beta_0 + \beta_1 \times x_1 + \beta_2 \times x_2 + \dots$$

In our case, that would become:

$$y = \beta_0 + \beta_1 \times \text{distance} + \beta_2 \times \text{hand}$$

(We'll investigate the effect of gender later.)

We want to estimate the parameters (β_0, β_1 , etc) from the data using maximum likelihood. The problem with this model is that y can take any value between $-\infty$ and $+\infty$, while the variable we are interested in, the probability of getting the sock in the box, p , has to be between 0 and 1. So we usually use a logistic link (or “logit link”) to connect p to y :

$$y = \log\left(\frac{p}{1-p}\right) \text{ and } p = \frac{e^y}{1+e^y} = \frac{1}{1+e^{-y}}$$

With this link, when $y = \infty, p = 1$, and when $y = -\infty, p = 0$. When $y = 0, p = 0.5$.

Now the expression $y = \beta_0 + \beta_1 \times \text{distance} + \beta_2 \times \text{hand}$ is called the **linear predictor** for p .

Analysis in Excel

We'll calculate y for each row in the spreadsheet, then p , the likelihood and the log(likelihood); put in the column headings as shown on the right.

	E	F	G	H	I
	hand.good	y	p	llh	log(llh)
1					
2					

Leave a blank column and put in some sensible starting values for the parameters. The Intercept determines y when Distance = Hand = 0. With Distance = 0, y (and p) will be quite high, so I put in 4 as the starting value ($y = 4$ means $p = 0.98$); y will go down as Distance increases, so I put -1 for the Distance parameter. I think Hand might make no difference, so I put 0.

	K	L
	Parameters	
	Intercept (B0)	4
	Distance (B1)	-1
	Hand (B2)	0

Calculate y : Next we translate the equation

$$y = \beta_0 + \beta_1 \times \text{distance} + \beta_2 \times \text{hand}$$

into Excel's code and put it in the first cell of the 'y' column:

$$= L2 + L3*D2 + L4*E2$$

We will copy this formula down the whole column, but we don't want L2, L3 and L4 to change, so we put in \$ signs. You can type in "\$" or place the cursor in the middle of (say) L2 and press F4. (If you aren't familiar with this, look for “absolute reference” in Excel help.)

$$= \$L\$2 + \$L\$3 * D2 + \$L\$4 * E2$$

Copy this down the whole column.

Calculate p : We translate the equation $p = \frac{1}{1+e^{-y}}$ and put it in the first cell of the 'p' column:

$$= 1 / (1 + EXP(-F2))$$

Copy this down the whole column; all the values should be between 0 and 1.

Calculate the likelihood contribution for each row : p is the probability of getting the sock in the box, ie. $result = 1$. The probability of $result = 0$ is $1 - p$. The likelihood is the probability of getting the result we actually observed, so we want p in the rows where $result = 1$ and $1 - p$ in the rows where $result = 0$. We can do this with the 'IF' function in Excel:

$$=IF(B2=1, G2, 1-G2)$$

Put this in cell H2 and copy down the column; check that you have the right values.

Calculate the log(likelihood) : Excel uses LN for natural logarithms, which is what we want (LOG means base 10 logarithms). So cell I2 has just:

$$= \text{LN}(H2)$$

Then we get the total log(likelihood) by adding up column I. I put the total in cell L10, below the parameter values:

$$=\text{SUM}(I:I)$$

N-pars	3
Log(llh)	-26.3592
AIC	58.71837

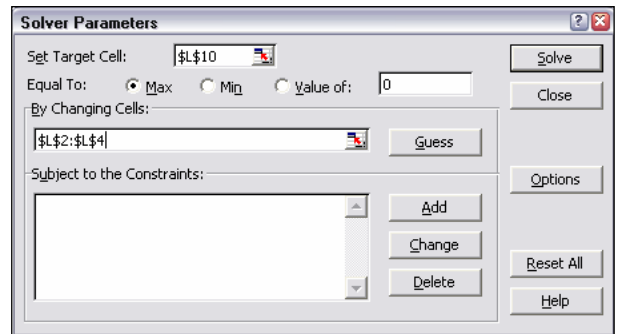
Save the spreadsheet.

Maximum Likelihood Estimation : We want to find the values for the parameters in cells L2 to L4 which give the maximum value for Log(llh).

Experiment with different values for the parameters and find values which produce bigger values of Log(llh) (remember that -24 is bigger than -26). When you get tired of that, use Solver to do it for you!

Open the Solver window (in Excel 2003, go to Tools > Solver; in Excel 2007, go to the Data tab and the Analysis group).

- Set Target Cell: $\$L\10 , or wherever you put the total log(likelihood); Solver uses absolute references, so if you type “L10” it will change it to “ $\$L\10 ”.
- We want Solver to find the ‘Equal to: Max’ value.
- By Changing Cells: $\$L\$2:\$L\4 , ie. the values of the parameters.



Click the “Solve” button and the Solver Result box opens. Check that ‘Solver found a solution...’, select ‘Keep Solver Solution’ and press ‘OK’.

Calculate AIC : We have run one model, which uses *distance* and *hand* as covariates. We’ll try other models and use AIC to compare them. So we need to calculate AIC. For this we use the formula:

$$\text{AIC} = -2 \times \text{Log}(\text{llh}) + 2 \times \text{Number of parameters}$$

I put the number of parameters (3 for this model) in cell L8, and calculated AIC in cell L12:

$$= -2 * L10 + 2 * L8$$

Write down the AIC for this model before you run the next one. Save the spreadsheet.

Running other models : First try a model without the *hand* covariate. You could modify the formula in column F to $= \$L\$2 + \$L\$3 * D2$, but it’s simpler just to ensure that the *hand* parameter is 0. To do this:

- Set the Hand (β_2) parameter in cell L4 to 0.
- Change the number of parameters in cell L8 to 2.
- Run Solver with “By Changing Cells:” set to “ $\$L\$2:\$L\3 ”.

How does the value of AIC compare with the previous model? Write it down before running the next model.

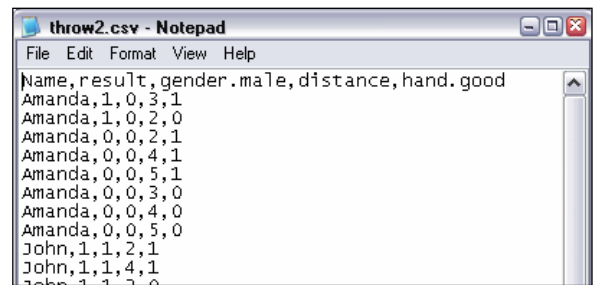
Run the model with *hand* but without *distance* (tip: use “ $\$L\$2, \$L\4 ”), and note the AIC. Also run the “null model”, which has no covariates, just the intercept. Compare AICs: which is the best model? Is there uncertainty about which model is best?

Analysis in R

Save the data as a .csv file : Make sure you have saved the Excel file with all the calculations. Then remove all the columns except for the original data (columns A to E). Now go to File > Save As..., change “Save as type:” to “(CSV) comma delimited”, give it a name like “trashball” and save it in the

Modelling folder. Click “Yes” in the confirmation box which pops up. Then close the file in Excel **without** saving it.

Find the file you just created, right click and chose “Open with...” Notepad. It should look like the screen shot right.



Import the data into R : Go to the Modelling folder and double-click on the **R** icon to start R. Go to File > Open script...” and open “Trashball.R”.

Follow the code in the script to import the data from the .csv file.

Take a look at the data and plot *result vs distance*. Since *result* only has values of 0 and 1, the dots will be on top of each other. One way to get around this is to add a small random value to *result*, which we can do with `jitter`.

Running the models : Logistic models are a type of Generalized Linear Model (GLM), which we can run with the `glm` function in R. Putting `family='binomial'` in the call to `glm` specifies a logistic model.

You specify the response and the covariates with the tilde (~), eg.

```
result ~ hand.good + distance
```

specifies a model with *result* as the response and *hand* and *distance* as covariates.

Run a model with *distance* as the covariate and look at the parameters. What do they mean?

Similarly, run models with *hand* and both *distance* and *hand* as covariates. Make sure you understand what the parameters mean in each model.

Also run a model with no covariates, called the “null model”; do that with `result ~ 1`. Then compare the value of AIC for all the models. Which is the best model? Is there uncertainty about which model is best?

The model matrix

The software packages PRESENCE and MARK use logistic models with the logit link. They don’t allow you to use ~ to specify your models, instead they use a “design matrix”. The `glm` function also uses a design matrix (called a “model matrix” in R), which you can display with `model.matrix`.

Take a look at the model matrix for the `result ~ hand.good + distance` model and compare it with the original data in Excel. Also compare with the model matrices for the other models you ran in R.

The `glm` function (like PRESENCE and MARK) uses matrix multiplication to calculate *y* from the model matrix and the vector of parameters. The script shows how this is done in R. The same method can be used to make predictions for new values of the covariates.

Key points

- With binary data, we want to estimate the probability of “success” – survival, detection, etc.
- If we have covariates (aka predictors or explanatory variables), we put these into a **linear predictor** and link this to the probability of success with a **link function**, usually the logistic (or “logit”) link.
- We can then obtain maximum likelihood estimates for the parameters of the model, calculate AIC and compare different models.

- In R we can specify models with the \sim convention; R converts this to a model matrix which it uses to calculate the linear predictor. In PRESENCE and MARK we work with the model matrix (aka design matrix).

Additional exercises

Add *gender* as a covariate to each of the models we ran above. Do this in Excel and in R. Does including gender improve the model? Is there evidence of a gender effect?

References

Morrell, C H; R E Auer. 2007. Trashball: A logistic regression classroom activity. *J Statistics Education* **15**:1. On-line at www.amstat.org/publications/jse/v15n1/morrell.html