

Design and Analysis of Distance Sampling Studies

Carl James Schwarz

Department of Statistics and Actuarial Science
Simon Fraser University
Burnaby, BC, Canada
cschwarz @ stat.sfu.ca

Part 4 Mark-recapture distance sampling

MRDS - Mark-Recapture Distance Sampling

WARNING: Tough sledding ahead!

Mark-Recapture Distance Sampling

- Allows $g(0) \neq 1$ by using capture-recapture to estimate detection probability.
 - $E[\hat{D}] = Dg(0)$ so bias can be considerable if $g(0) < 1$.
- Usually has two observers who can (independently) identify individual animals and match animals seen by both, seen by one and not the other. This gives an estimate of animals missed and hence $g(0)$.
- 'Observer' is a generic term – any two methods of observation along the transect will work.

Implemented in DISTANCE through a call to *R*, but not yet fully featured.

CHECK TO SEE THAT *R* is installed and working.

Types of observer data (which influences analysis)

- Independent configuration. Two observers record location and distance of animals without communication with each other. Treatment of observers is symmetric.
- Trial configuration. Observer 2 “seeds” animals which are then detected by Observer 1. E.g. Observer 2 could be radio detections. Observers are not treated symmetrically.
- Removal configuration. Observer 1 “removes” animals and Observer 2 only looks for new animals. E.g. Observer 1 could be map of known nests seen in previous surveys.

Removal configurations not yet implemented in DISTANCE.

MRDS - Golf Tee Example

MRDS data structure

- Usual fields as before +
- **NEW**: Unique identifier for each object and 2 lines for each object (one for each observer)
- **NEW**: Field identifying the observer, the distance, and if each observer detected (0=no, 1=yes)

Refer to GolfTee example. **Independent configuration** of observers.

- There is a field for *sex* (I didn't know that golf tees had sexes) which actually is color (green=0 or yellow=1).
- There is a field for *exposure* - if tees are above grass (exposed=1) or within grass (exposed=0).
- Two teams = two observers.

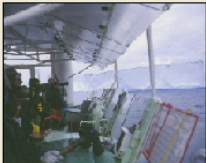
Notice that there is only 1 transect (not a good idea!).

MRDS - Golf Tee Example

Import the data in the usual fashion.

New Project Setup Wizard - Step 3: Survey Methods

In this screen, you tell Distance about your survey methods. Click 'Help' to find out more about each option.



*Minke whale line transect surveys, Antarctic Ocean
Photo: Peter Corkeron*

Type of survey

- ☒ Line transect
- ☐ Point transect
- ☐ Cue count

Observer configuration

- ☐ Single observer
- ☒ Double observer

Distance measurements

- ☒ Perpendicular distance
- ☐ Radial distance and angle

Observations

- ☒ Single objects
- ☐ Clusters of objects

Sampling fraction
This option has been moved to the Multipliers page.

Help Cancel < Back Next > Finish

MRDS - Golf Tee Example

... Import the data in the usual fashion - lengths in m, and area in m^2 .

New Project Setup Wizard - Step 4: Measurement Units

Please specify the measurement units for your data.


If you want to analyze the data using different units, you can do so after completing this wizard (in the Units tab of the Data Filter). Click 'Help' for more information.

Units of original measurements


Distance:

Transect:

Area:



An aircraft ideally suited to aerial line transects
Photo: John Reinhardt



Distance intervals marked on wing struts
Photo: Rich Guenzel

MRDS - Golf Tee Example

... Import the data in the usual fashion - lengths in m, and area in m^2 .

New Project Setup Wizard - Step 4: Measurement Units

Please specify the measurement units for your data.


If you want to analyze the data using different units, you can do so after completing this wizard (in the Units tab of the Data Filter). Click 'Help' for more information.

Units of original measurements


Distance:

Transect:

Area:



An aircraft ideally suited to aerial line transects
Photo: John Reinhardt



Distance intervals marked on wing struts
Photo: Rich Guenzel

MRDS - Golf Tee Example

Create a *Data Filter* in the usual way, truncating at 4 m.

Data Filter Properties: [Truncation 4 meters]

Data selection | Intervals | **Truncation** | Units

Truncation of exact distance measurements

Right truncation

- ☐ Right truncate at largest observed distance
- ☐ Discard the largest percent of distances
- ☒ Discard all observations beyond

Left truncation

- ☒ No left truncation
- ☐ Discard all observations within

Truncation for cluster size estimation (where required)

Right truncation

- ☒ Same as that specified above
- ☐ Discard all observations beyond

Defaults | Name: | |

There are four detection functions!

- Observer 1 detection function $p_1(y, z)$ where z are covariates
- Observer 2 detection function $p_2(y, z)$
- Observer 1 detects — Observer 2 detects $p_{1|2}(y, z)$
- Observer 2 detects — Observer 1 detects $p_{2|1}(y, z)$

What is the shape of detection functions if Observer 1 is more experienced than Observer 2?

What is the shape of detection functions if Observer 1 and Observer 2 are both equally experienced.

The four detection functions!

- Observer 1 detection function $p_1(y, z)$ where z are covariates
- Observer 2 detection function $p_2(y, z)$
- Observer 1 detects — Observer 2 detects $p_{1|2}(y, z)$
- Observer 2 detects — Observer 1 detects $p_{2|1}(y, z)$

What is the shape of detection functions if Observer 1 is more experienced than Observer 2?

What is the shape of detection functions if Observer 1 and Observer 2 are both equally experienced.

Simplest (unreasonable) Case

Suppose that all animals were equally detectable at all distances, and observers are independent, but a different detection probability for each observer?

What do the detection functions look like?

Then we get for each animal:

- $P(1, 0) = p_1(1 - p_2)$
- $P(0, 1) = (1 - p_1)p_2$
- $P(1, 1) = p_1p_2$

Now a standard mark-recapture experiment and $\hat{N} = \frac{n_1 n_2}{n_{12}}$, i.e. Petersen estimator

More complex (but still unreasonable) Case

Suppose that detectability of animals varies by distance, and observers are independent, but a different detection probability for each observer (that varies by distance)?

What do the detection functions look like?

Then we get for each animal:

- $P(1, 0)(y) = p_1(y) (1 - p_2(y))$
- $P(0, 1)(y) = (1 - p_1(y)) p_2(y)$
- $P(1, 1)(y) = p_1(y) p_2(y)$

Break distances into bands, do a separate Petersen in each band, and add them up.

Reality Case

Suppose that detectability of animals varies by distance, and observers are NOT independent, but a different detection probability for each observer (that varies by distance)?

What do the detection functions look like?

Then we get for each animal:

- $P(1, 0)(y) = p_1(y) (1 - p_{2|1}(y))$
- $P(0, 1)(y) = (1 - p_{1|2}(y)) p_2(y)$
- $P(1, 1)(y) = p_1(y)p_{2|1}(y) = p_{1|2}(y)p_2(y)$

Now we have to estimate all 4 detection functions!

Observers acting independently does NOT imply that detection functions are independent.

Suppose detection depends on distance y and cluster size s with perfect detection at the origin. If observer 1 detects an object at distance y it is more likely to be a larger cluster which implies that observer 2 will detect it.

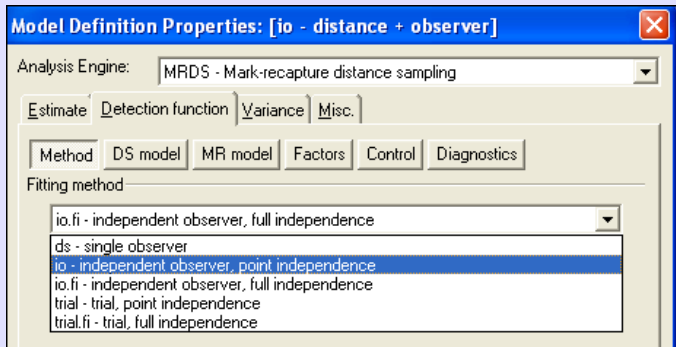
We call this *Point Independence* (independence only at $y = 0$) to distinguish it from *Complete Independence*

MRDS - Golf Tee Example

Specifying models in DISTANCE in MRDS.

Four types of models depending on configuration and type of independence assumed:

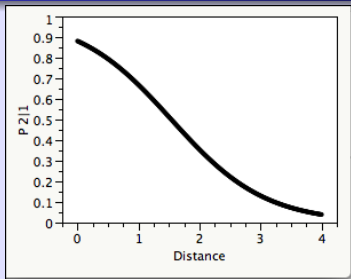
- independent configuration, point independence
- independent configuration, full independence
- trial configuration, point independence
- trail configuration, full independence



Two different models for detection must be specified:

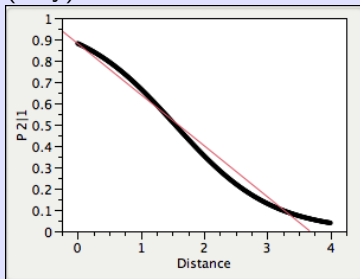
- Distance Sampling (DS) model ($p_1(y, z)$ and $p_2(y, z)$).
- Conditional probability of detection models (MR) ($p_{1|2}(y, z)$ and $p_{2|1}(y, z)$)

MRDS - Golf Tee Example



Develop a model for this curve?

Try a straight line, $P_{2|1} = \beta_0 + \beta_1(\text{distance})$, but not satisfactory (why)



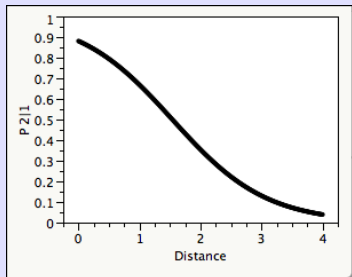
MRDS - Golf Tee Example

To model these 'S'-shaped curves and prevent the line from going > 1 or < 0 , we often model

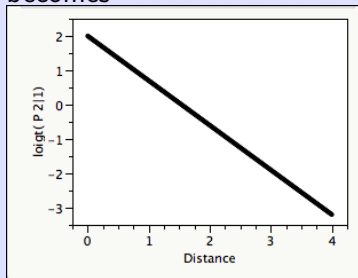
$$\text{logit}(p) = \ln \left(\frac{p}{1-p} \right)$$

Probability	Odds	Logit
.01	1:99	-4.59
.1	1:9	-2.20
.5	1:1	0
.6 6:4 or 3:2 or 1.5		.41
.9	9:1	2.20
.99	99:1	4.59

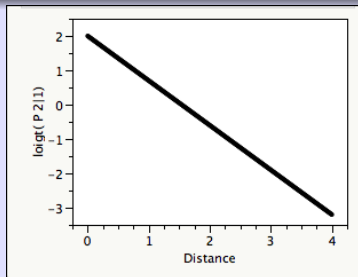
MRDS - Golf Tee Example



becomes



MRDS - Golf Tee Example



Now modeling on logit-scale

$$\text{logit}(p_{2|1}) = \beta_0^* + \beta_1^*(\text{distance})$$

keeps $0 < p < 1$.

The reverse transformation can be written in two ways:

$$p = \frac{e^{\text{log-odds}}}{1 + e^{\text{log-odds}}} = \frac{1}{1 + e^{-\text{log-odds}}}$$

MRDS - Golf Tee Example

Specifying a model for the MR component in DISTANCE

Need to use a generalized linear model syntax for the right hand part of the model.

Can use continuous and categorical (factors) in the model.

Examples:

- $\text{logit}(p_{2|1}) = 1$ corresponds to $\text{logit} = \beta_0$.
- $\text{logit}(p_{2|1}) = \text{distance}$ corresponds to $\text{logit} = \beta_0 + \beta_1(\text{distance})$
- $\text{logit}(p_{2|1}) = \text{sex}$ corresponds to $\text{logit} = \beta_0 + \beta_1(I(\text{sex} = f))$ where $I()$ is an indicator function
- $\text{logit}(p_{2|1}) = \text{distance} + \text{sex}$ corresponds to $\text{logit} = \beta_0 + \beta_1(\text{distance}) + \beta_2(I(\text{sex} = f))$
- $\text{logit}(p_{2|1}) = \text{distance} + \text{sex} + \text{distance} : \text{sex}$ corresponds to $\text{logit} = \beta_0 + \beta_1(\text{distance}) + \beta_2(I(\text{sex} = f)) + \beta_3(I(\text{sex} = f))(\text{distance})$

Draw the curves (in logit space) for each of these.

MRDS - Golf Tee Example

Creating MR component in MRDS

Define the component and independence structure. We will start with independent configuration and full independence among observers.

Model Definition Properties: [io - distance + observer]

Analysis Engine:

Fitting method

Notice poor choice of words for “io” option for configuration.

MRDS - Golf Tee Example

Creating MR component in MRDS

Select the MR tab:

The screenshot shows the 'Model Definition Properties' dialog box for the 'MRDS - Mark-recapture distance sampling' analysis engine. The 'Detection function' tab is selected. Within this tab, the 'MR model' sub-tab is highlighted with a red rectangle. The 'Mark-recapture model' section contains a text box with the formula 'distance'. Below this, the 'Class of model' is set to 'Generalized linear model' (selected with a radio button). The 'Link function' is set to 'logit' in a dropdown menu, and the 'Model formula' is 'distance'. A red rectangle highlights the 'Link function' dropdown and the 'Model formula' text box. At the bottom, the 'Name' field is filled with 'jo.fi DS() MR(d)'. Buttons for 'Defaults', 'OK', and 'Cancel' are at the bottom right.

Model Definition Properties: [jo - distance + observer]

Analysis Engine: MRDS - Mark-recapture distance sampling

Estimate Detection function **Variance** Misc.

Method DS model **MR model** Factors Control Diagnostics

Mark-recapture model

This is the model for probability of detection by a single observer, p_{ij} where $j=1$ or 2 . Note that distance is only a covariate if it is included in the model formula.

Class of model: ☒ Generalized linear model
☐ Generalized additive model

Link function: logit

Model formula: distance
(Linear/additive predictor)

Defaults Name: jo.fi DS() MR(d) OK Cancel

CAUTION: Variable names are a bit trick – see later slides.

MRDS - Golf Tee Example

Creating MR component in MRDS

List categorical (factor) variables in FACTORS tab.

The screenshot shows the 'Model Definition Properties' dialog box for the 'io - distance + observer' model. The 'Analysis Engine' is set to 'MRDS - Mark-recapture distance sampling'. The 'Factors' tab is selected and highlighted with a black box. Below the tabs, the 'Factor definition' section contains the instruction: 'Here, you list variables in the DS and MD formulae that should be treated as factors. Separate each variable name with a comma.' The 'Factors:' text box contains the text 'observer, sex, exposed', which is also highlighted with a black box. Other tabs visible include 'Estimate', 'Detection function', 'Variance', 'Misc.', 'Method', 'DS model', 'MR model', 'Control', and 'Diagnostics'.

CAUTION: Variable names are a bit trick – see later slides.

No harm in listing all variables in dataset here even if not used in model.

MRDS - Golf Tee Example

Creating DS component in MRDS

Switch to DS Tab:

The screenshot shows a software window titled "Model Definition Properties: [io - distance + observer]". Inside, the "Analysis Engine" is set to "MRDS - Mark-recapture distance sampling". Below this are four tabs: "Estimate", "Detection function", "Variance", and "Misc.". Under the "Detection function" tab, there are six sub-tabs: "Method", "DS model", "MR model", "Factors", "Control", and "Diagnostics". The "DS model" sub-tab is currently selected. Below these tabs, the text reads: "Distance sampling model" followed by a horizontal line and then "This model is not used in the current fitting method. To change the fitting method, click on the 'Method' button above."

Why? If fully independent, then MR functions provides ALL information needed.

MRDS - Golf Tee Example

Combination of DS and MR components needed:

Fitting method	Configuration	Independence	DS model	MR model
ds	single ¹	-	yes	no
io	independent	point	yes	yes
io.fi	independent	full	no	yes
trial	trial	point	yes	yes
trial.fi	trial	full	no	yes
removal ²	removal	point	yes	yes
removal.fi ²	removal	full	no	yes

MRDS - Golf Tee Example

io.fi Data(< 4) DS() MR(dis): Fit Model ...

Analysis 2: [io.full d+o+d_0] Set: [Set 1]

Analysis

Name:

Created: 11/17/2012 11:14:47 PM

Run:

Survey

Data filter

1 Default Data Filter	<input type="button" value="Properties ..."/> <input type="button" value="New ..."/>
2 Data <4 m	

Model definition

1 Default Model Definition	<input type="button" value="Properties ..."/> <input type="button" value="New ..."/>
2 io	
3 io.fi DS() MR(d)	

Comments

Inputs

Log

Results

MRDS - Golf Tee Example

io.fi Data(< 4) DS() MR(dis): ... Results ...

Here is a list of the variable (field) names you can use in the model. See reference manual for rules.

The following fields will be written to the data file, and formulae. Note that you should use the new names, not the old formulae, and that formulae names are case sensitive.

Format: [layer name].[field name] AS new name

[Observation].[Perp distance] AS distance

[Observation].[object] AS object

[Observation].[observer] AS observer

[Observation].[detected] AS detected

[Observation].[Cluster Size] AS cluster.size

[Observation].[Sex] AS sex

[Observation].[Exposure] AS exposure

[Line transect].[Label] AS label

[Line transect].[Line length] AS line.length

[Region].[Label] AS stratum.label

[Region].[Area] AS area

[Study area].[Label] AS global.label

MRDS - Golf Tee Example

io.fi Data(< 4) DS() MR(dis): ... Results ...

Summary for io.fi object

```
Number of observations : 162
Number seen by primary : 124
Number seen by secondary : 142
Number seen by both : 104
AIC : 701.3888
```

Conditional detection function parameters:

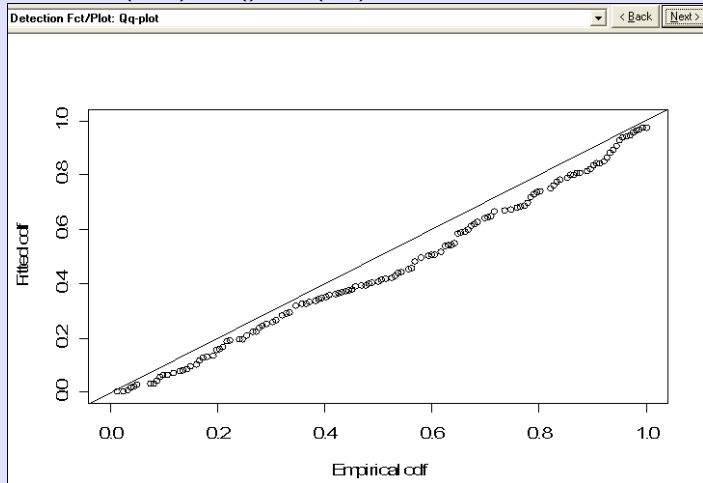
	estimate	se
(Intercept)	2.8806278	0.3439334
distance	-0.9179038	0.1478512

	Estimate	SE	CV
Average p	0.8705244	0.024878308	0.028578531
Average primary p(0)	0.9468804	0.015767751	0.016652315
Average secondary p(0)	0.9468804	0.015767751	0.016652315
Average combined p(0)	0.9971783	0.002726864	0.002734580
N in covered region	186.0947321	7.480834001	0.040199064

What does this model say?

MRDS - Golf Tee Example

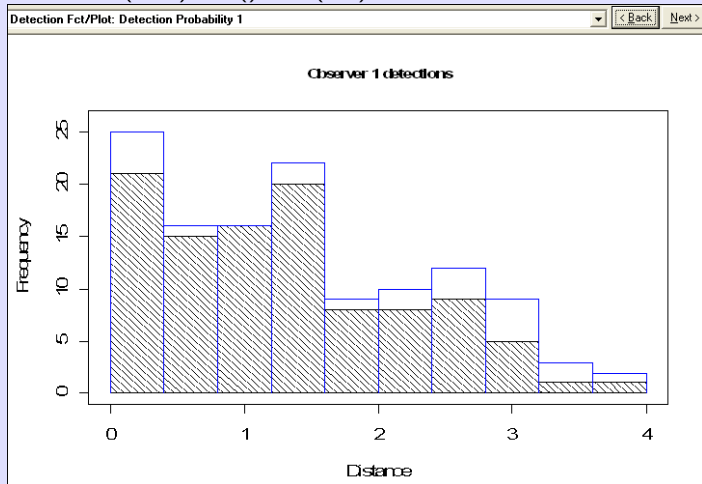
io.fi Data(< 4) DS() MR(dis): ... Results ...



What does this model say?

MRDS - Golf Tee Example

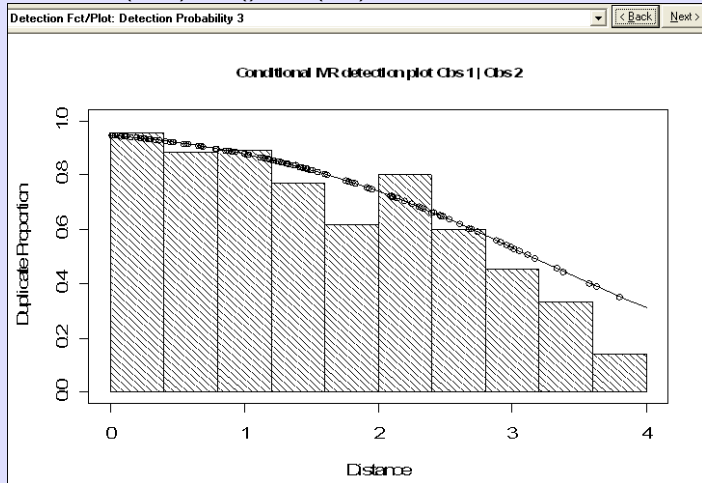
io.fi Data(< 4) DS() MR(dis): ... Results ...



What does QQ-plot indicate?

MRDS - Golf Tee Example

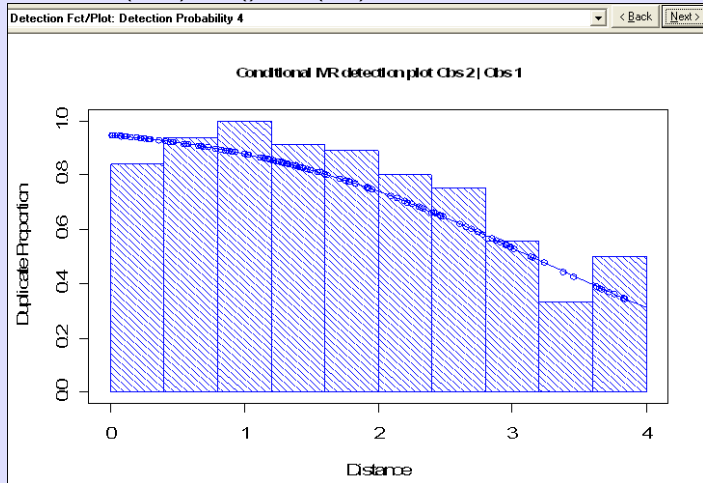
io.fi Data(< 4) DS() MR(dis): ... Results ...



What does this and next plot say?

MRDS - Golf Tee Example

io.fi Data(< 4) DS() MR(dis): ... Results ...



What does this and previous plot say?

MRDS - Golf Tee Example

io.fi Data(< 4) DS() MR(dis): ... Results ...

Density Estimates and associated quantities

Summary statistics:

Region	Area Covered	Area	Effort	n	k	ER	se.ER	cv.ER	
1	1	1680	1680	210	162	1	0.7714286	NaN	NaN

Abundance:

Label	Estimate	se	cv	lcl	ucl	df
1 Total	186.0947	14.64169	0.07867871	105.2985	328.8866	1.326888

Density:

Label	Estimate	se	cv	lcl	ucl	df
1 Total	0.1107707	0.008715294	0.07867871	0.06267765	0.1957658	1.326888

Notice that because there was only 1 transect, variance is understated.

MRDS - Golf Tee Example

Fit the `io.fi Data(< 4) DS() MR(dis+obs)` model.

Don't forget to specify `observer` as a factor (category). What do you conclude?

MRDS - Golf Tee Example

io.fi Data(< 4) DS() MR(dis+obs): ... Results ...

Conditional detection function parameters:

	estimate	se
(Intercept)	2.6103357	0.3627208
distance	-0.9179028	0.1503738
observer2	0.6418561	0.3132827

	Estimate	SE	CV
Average p	0.8797702	0.025122625	0.028555895
Average primary p(0)	0.9315238	0.020911540	0.022448744
Average secondary p(0)	0.9627518	0.013478134	0.013999594
Average combined p(0)	0.9974494	0.003511698	0.003520678
N in covered region	184.1390000	7.267307728	0.039466423

What does this model look like?

MRDS - Golf Tee Example

io.fi Data(< 4) DS() MR(dis+obs): ... Results ...

What do you conclude?

MRDS - Golf Tee Example

Fit the `io.fi Data(< 4) DS() MR(dis+obs+dist:obs)` model.

Don't forget to specify observer as a factor (category).

Interaction denoted by colon in model formulae.

What do you conclude?

Fit the

```
io.fi Data(< 4) DS() MR(dis+obs+dist:obs+color+exposure)
```

Don't forget to specify observer, color, exposure as a factor (category).

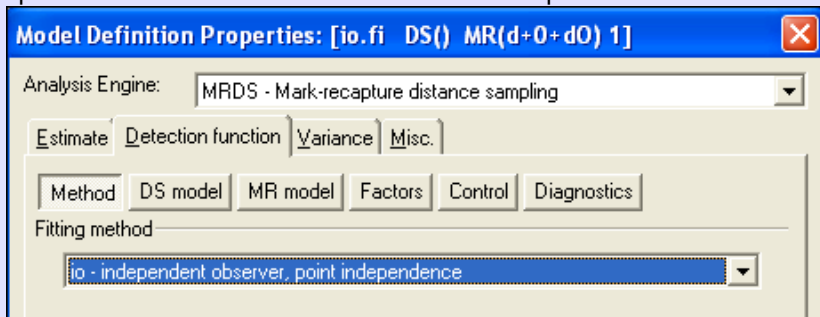
What do you conclude?

MRDS - Golf Tee Example

Point independence.

Here you assume observers have independent detections only a ONE point (typically the transect line).

This now requires a separate model for the detection function specified in a similar fashion as the MR component.



The screenshot shows the 'Model Definition Properties' dialog box. The title bar reads 'Model Definition Properties: [io.fi DS() MR(d+0+d0) 1]'. The 'Analysis Engine' is set to 'MRDS - Mark-recapture distance sampling'. The 'Detection function' tab is selected. Within this tab, the 'Method' sub-tab is active, showing a list of fitting methods. The method 'io - independent observer, point independence' is selected. Other tabs include 'Estimate', 'Variance', 'Misc.', 'DS model', 'MR model', 'Factors', 'Control', and 'Diagnostics'.

Model Definition Properties: [io.fi DS() MR(d+0+d0) 1]

Analysis Engine: MRDS - Mark-recapture distance sampling

Estimate Detection function Variance Misc.

Method DS model MR model Factors Control Diagnostics

Fitting method

io - independent observer, point independence

MRDS - Golf Tee Example

Fit the

io.pi Data(< 4) DS(HN)

MR(dis+obs+dist:obs+color+exposure)

Model Definition Properties: [io.fi DS() MR(d+0+d0) 1]

Analysis Engine: MRDS - Mark-recapture distance sampling

Estimate Detection function Variance Misc.

Method DS model MR model Factors Control Diagnostics

Distance sampling model

This is the model for probability of detection by one or more observers at a given distance, $p.(distance)$

Key function: half normal

Model for scale parameter of key function

☒ Scale parameter is a constant (CDS)

☐ Scale parameter is a function of additional covariates (MCDS)

Formula:

MRDS - Golf Tee Example

Fit the io.pi Data(< 4) DS(HN) MR(dis)

Analysis 1: [New Analysis] Set: [Set 1]

Analysis

Name:

Created: 11/17/2012 8:45:38 PM

Run:

Survey

Data filter

1 Default Data Filter
2 Data < 4 m

Model definition

2 io
3 io.fi DS() MR(d)
4 io.fi DS() MR(d+0)
5 io.fi DS() MR(d+0+d0)
6 io.pi DS(HN) MR(d)

Comments

Inputs

Log

Results

MRDS - Golf Tee Example

Model: $\text{io.pi Data}(< 4) \text{ DS(HN) MR(dis)}$

What do you conclude?

MRDS - Golf Tee Example

More complex models(!)

$p_{j 3-j}(y, z)$	PI	FI	
	$g_j(y, z)$	ΔAIC	\hat{N}
1 <i>D</i>	1	64.2	232.0
2 <i>D</i>	<i>C</i>	59.6	236.9
3 <i>D</i>	<i>S</i>	66.2	232.0
4 <i>D</i>	<i>E</i>	66.2	232.0
5 <i>D</i>	<i>C + E</i>	60.7	237.0
6 <i>D</i>	<i>C + S</i>	60.5	236.0
7 <i>D + P</i>	<i>C</i>	55.9	236.8
8 <i>D + C</i>	<i>C</i>	54.4	237.0
9 <i>D + S</i>	<i>C</i>	59.2	236.9
10 <i>D + E</i>	<i>C</i>	25.4	237.5
11 <i>D + C + P</i>	<i>C</i>	50.7	236.9
12 <i>D + S + P</i>	<i>C</i>	55.5	236.9
13 <i>D + E + P</i>	<i>C</i>	21.8	237.4
14 <i>D + C + P + E</i>	<i>C</i>	6.7	238.6
15 <i>D + S + P + E</i>	<i>C</i>	19.9	237.5
16 <i>D + S + P + C</i>	<i>C</i>	49.8	237.0
17 <i>D + P + C + S + E</i>	<i>C</i>	4.3	239.0
18 <i>D + P + C + S + E + C : E</i>	<i>C</i>	2.8	250.8
19 <i>D + P + C + S + E + S : E</i>	<i>C</i>	6.2	239.1
20 <i>D + P + C + S + E + S : C</i>	<i>C</i>	1.7	239.0
21 <i>D + P + C + S + E + S : C + C : E</i>	<i>C</i>	0.0	252.0
22 <i>D + P + C + S + E + S : C + S : E</i>	<i>C</i>	2.7	240.7
23 <i>D + P + C + S + E + E : C + S : E</i>	<i>C</i>	4.7	251.2
24 <i>D + P + C + S + E + E : C + S : E + C : S</i>	<i>C</i>	0.7	271.8

P=Obs; C=Color; S=Size; E=Exposure

Harbour Porpoise sample data from 1994 SCANS survey

- Independent configuration of observers.
- 98 transect lines, 1 km half-width of varying lengths (km)
- Recorded distance to group, group size, sex (groups are composed of a single sex), and exposure (a detectability factor related to weather and other attributes)
- Total survey region is 889600 km²

Data set is large, so model take a few seconds to fit!

Deals with $g(0) \neq 1$.

Survey Protocol

- Independent configuration.
- Trial configuration.
- Removal configuration (not yet implemented in MRDS).

Data structure:

- The usual + object identifier +
- Field for Observer (limit currently is 2) +
Detected (1=yes, 0=no)

Group the data together in the usual way. Import in the usual way.

Model Building.

- Start with small models and build up to more complex models
 - Full independence vs. Point independence
 - Few covariates vs. many covariates
 - Complex models lead to cases where categories of animals seen by one observer are never seen by the other observer leading to an infinite population size!
- DS (HN or HR, but no series); MR (logistic regression)
- Use Linear Model syntax, e.g. distance + sex + distance:sex
- AIC for model comparison over all options within same data filter

R package has more features, but less easy to use.