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!

MRDS - Introduction

Mark-Recapture Distance Sampling

- Allows $g(0) \neq 1$ by using capture-recapture to estimate detection probability.
  - $E[\widehat{D}] = Dg(0)$ so bias can be considerable if $g(0) < 1$.
- Usually has two observers who can (independently) identity 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.

MRDS - Introduction

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)

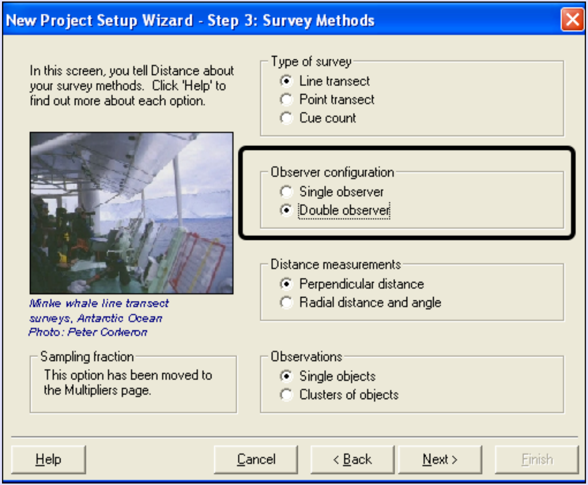
Refer to GolfTee example. **Independent configuration** of observers.

- 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!).

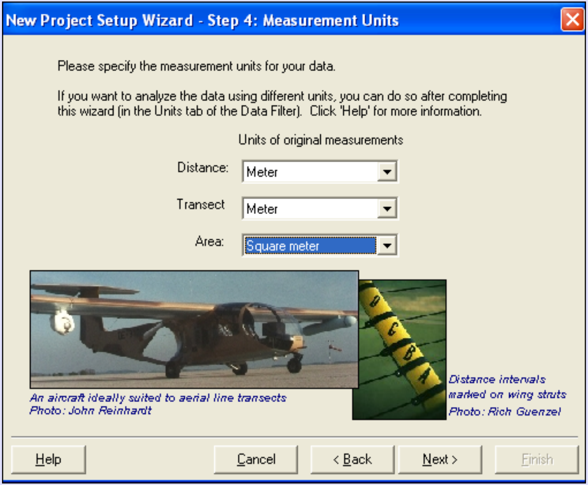
MRDS - Golf Tee Example

Import the data in the usual fashion.

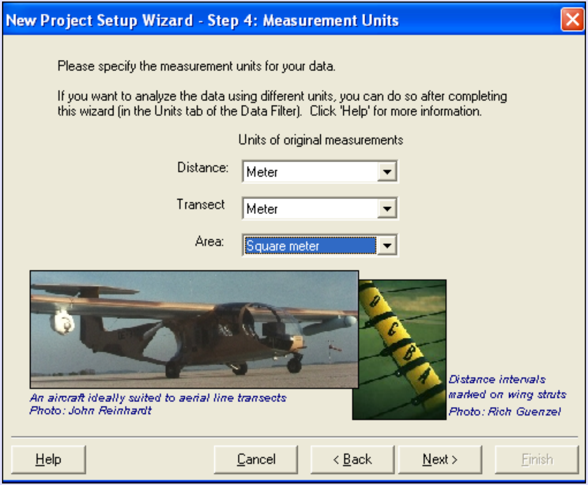


MRDS - Golf Tee Example

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

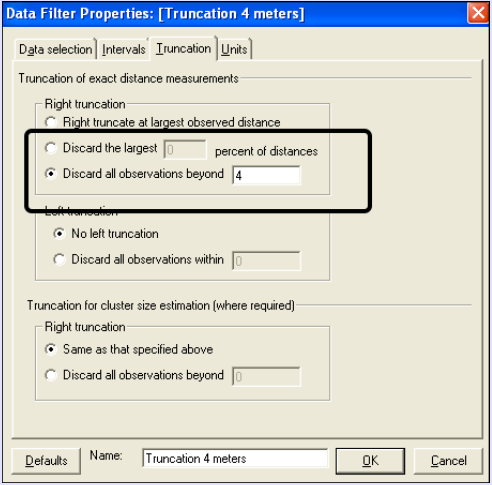


. . . Import the data in the usual fashion - lengths in m, and area in m^2.



MRDS - Golf Tee Example

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



MRDS - Golf Tee Example

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.

MRDS - Golf Tee Example

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.

MRDS - Golf Tee Example

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:

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

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

MRDS - Golf Tee Example

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:

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

Now we have to estimate all 4 detection functions!

MRDS - Golf Tee Example

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

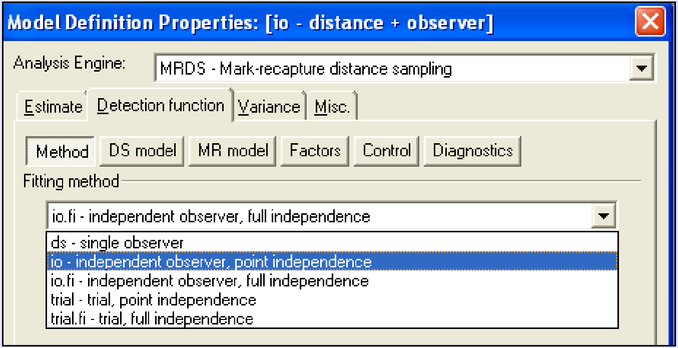
Suppose detection depends on distance y and cluster size s with perfect detection at the origin. If observer 1 detects an object at distance y it is more likely to be a larger cluster which implies that observer 2 will detect it.

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

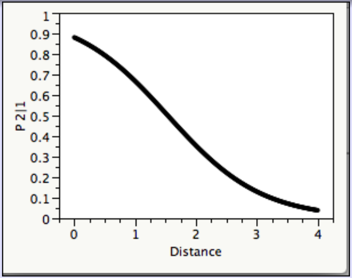


MRDS - Golf Tee Example

Two different models for detection must be specified:

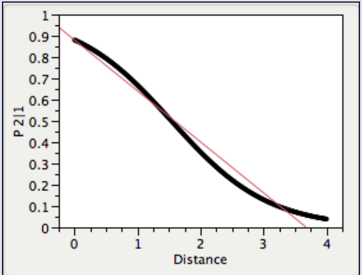
- Distance Sampling (DS) model ($p_1(y, z)$ and $p_2(y, z)$.
- Conditional probability of detection models (MR) ($p_{1|2}(y, z)$ and $p_{2|1}(y, z)$

MRDS - Golf Tee Example



Develop a model for this curve?

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

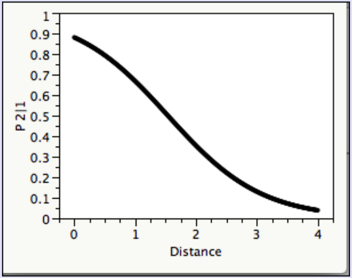


MRDS - Golf Tee Example

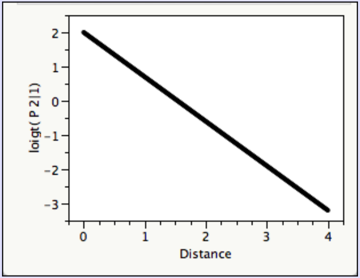
To model these 'S'-shaped curves and prevent the line from going > 1 or < 0, we often model

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

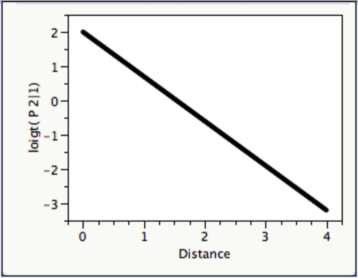
Probability	Odds	Logit
.01	1:99	-4.59
.1	1:9	-2.20
.5	1:1	0
.6	6:4 or 3:2 or 1.5	.41
.9	9:1	2.20
.99	99:1	4.59


P 2|1
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
1
2
3
4
Distance

becomes


loigt(P 2|1)
2
1
0
-1
-2
-3
0
1
2
3
4
Distance

MRDS - Golf Tee Example



Now modeling on logit-scale

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

keeps $0 < p < 1$.

The reverse transformation can be written in two ways:

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

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:

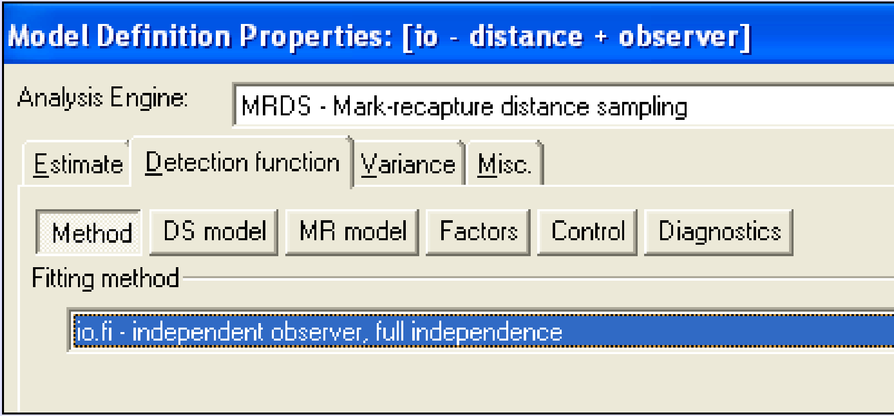
- $logit(p_{2|1}) = 1$ corresponds to $logit = \beta_0$.
- $logit(p_{2|1}) = distance$ corresponds to $logit = \beta_0 + \beta_1(distance)$
- $logit(p_{2|1}) = sex$ corresponds to $logit = \beta_0 + \beta_1(I(sex = f))$ where $I()$ is an indicator function
- $logit(p_{2|1}) = distance + sex$ corresponds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

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.

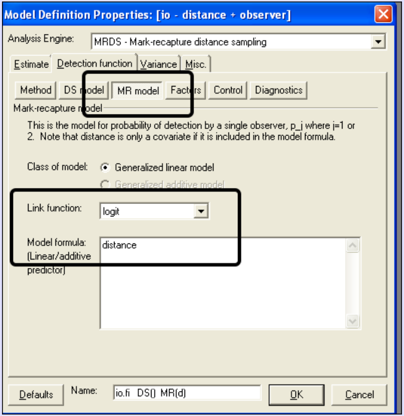


Notice poor choice of words for “io” option for configuration.

MRDS - Golf Tee Example

Creating MR component in MRDS

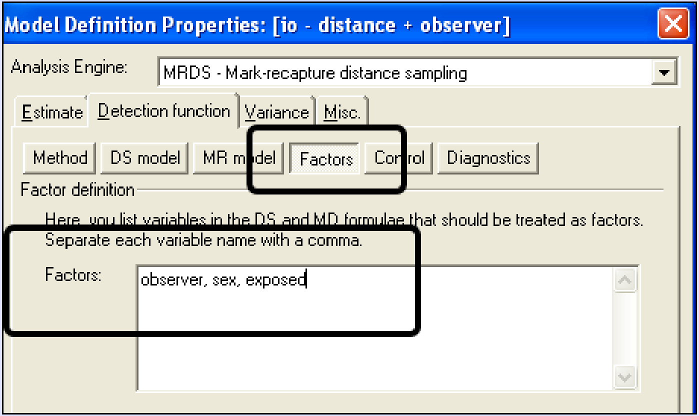
Select the MR tab:



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

MRDS - Golf Tee Example

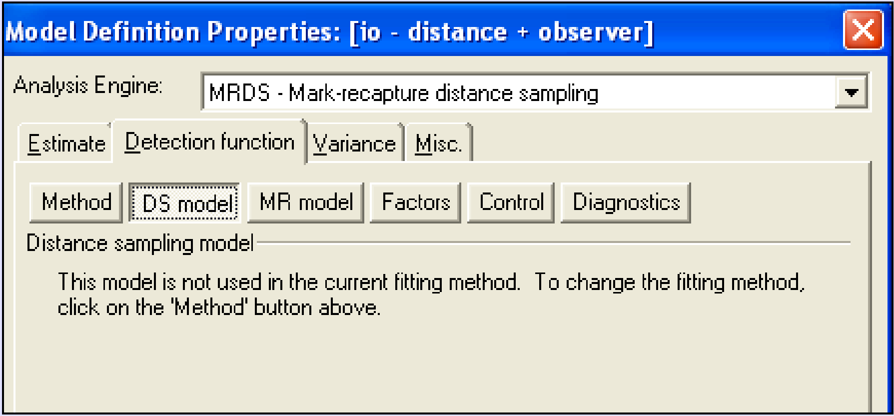
Creating MR component in MRDS
List categorical (factor) variables in FACTORS tab.



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:



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

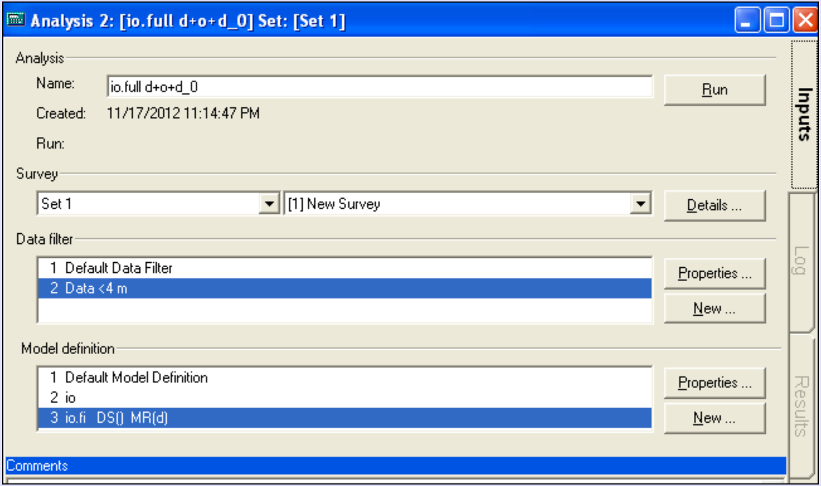
MRDS - Golf Tee Example

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 ...



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 o
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

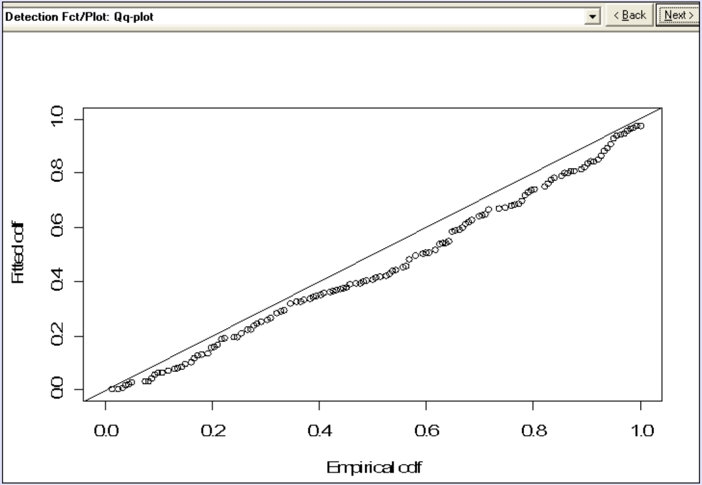
io.fi Data(< 4) DS() MR(dis): ... Results ...

```
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

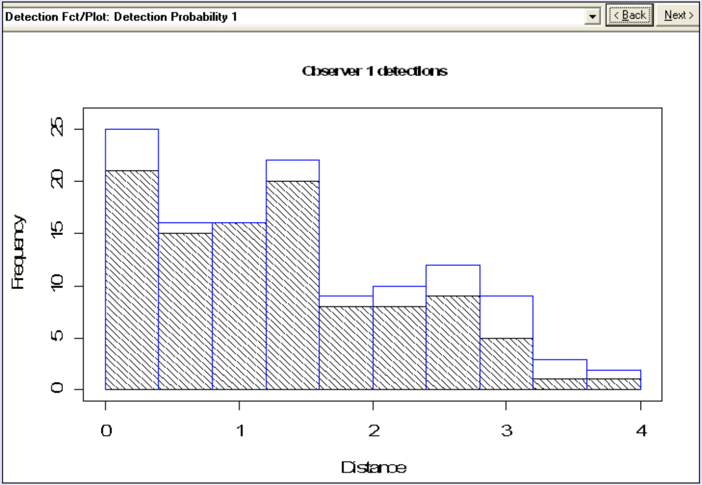
io.fi Data(< 4) DS() MR(dis): . . . Results . . .



What does this model say?

MRDS - Golf Tee Example

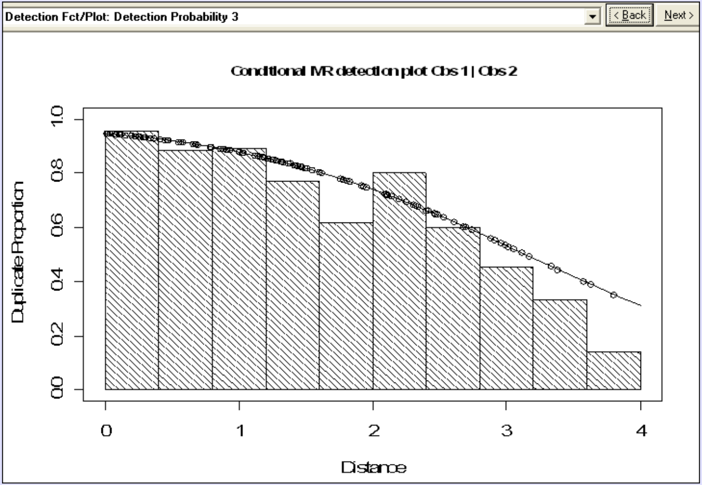
io.fi Data(< 4) DS() MR(dis): ... Results ...



What does QQ-plot indicate?

MRDS - Golf Tee Example

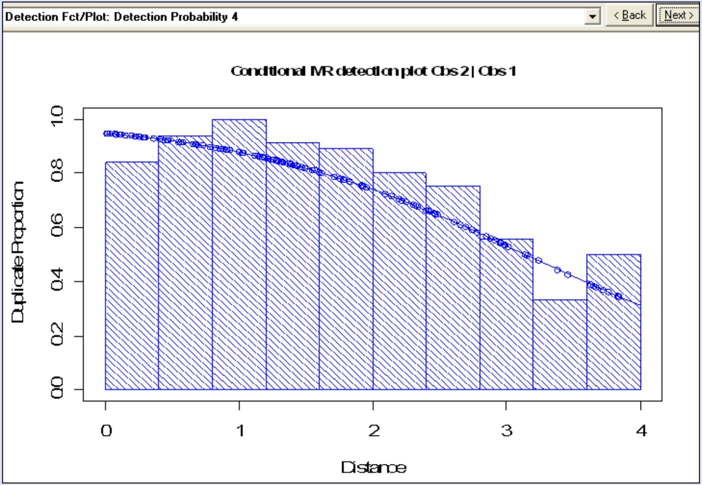
io.fi Data(< 4) DS() MR(dis): ... Results ...



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 CoveredArea 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.

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?

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).

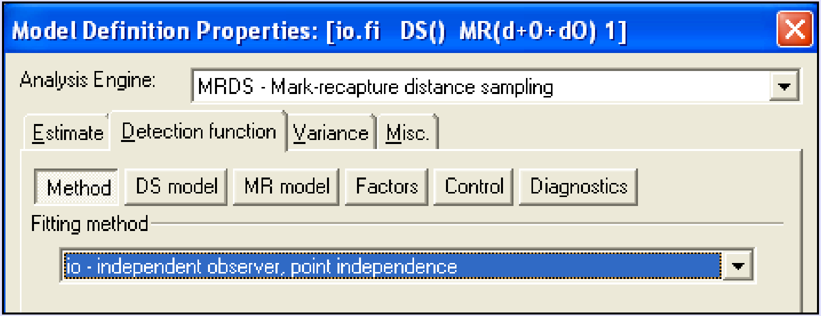
What do you conclude?

MRDS - Golf Tee Example

Point independence.

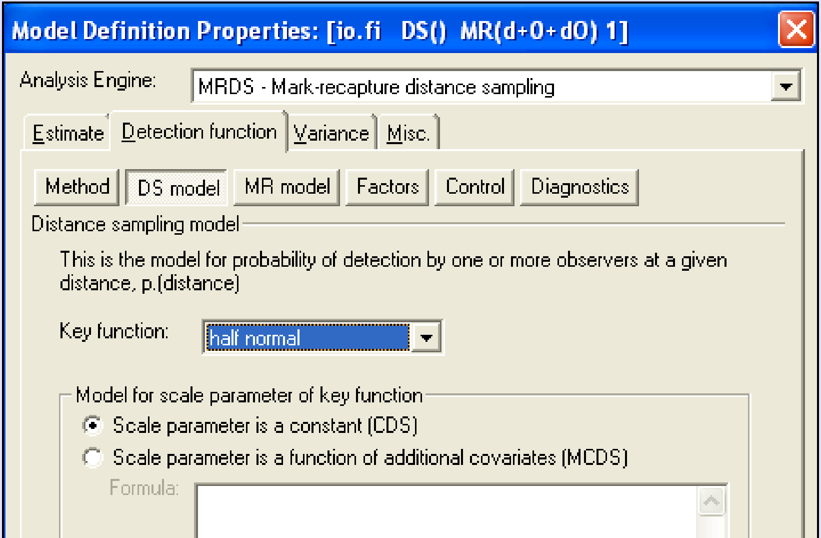
Here you assume observers have independent detections only a ONE point (typically the transect line).

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

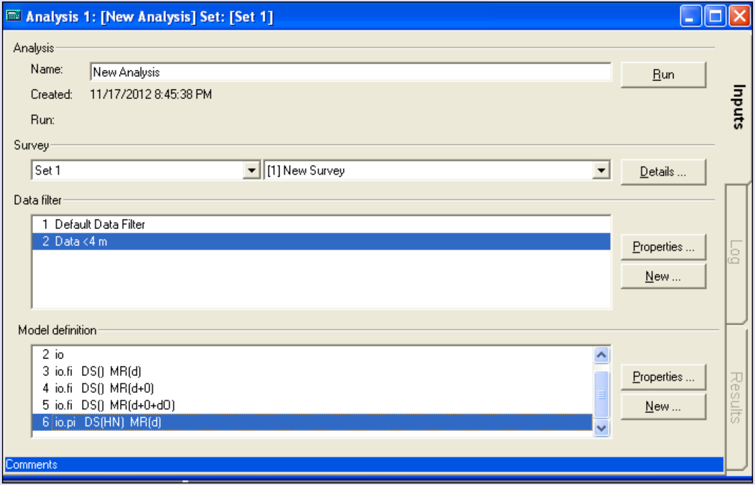


MRDS - Golf Tee Example

Fit the
io.pi Data($<$ 4) DS(HN)
MR(dis+obs+dist:obs+color+exposure)



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



MRDS - Golf Tee Example

Model: io.pi Data(< 4) DS(HN) MR(dis)

What do you conclude?

MRDS - Golf Tee Example

More complex models(!)

	$p_{j\mid 3-j}(y,\underline{z})$	PI			FI
		$g_{\cdot}(y,\underline{z})$	ΔAIC	$\widehat{N}$	ΔAIC
1	D	1	64.2	232.0	67.6
2	D	C	59.6	236.9	
3	D	S	66.2	232.0	
4	D	E	66.2	232.0	
5	D	$C+E$	60.7	237.0	
6	D	$C+S$	60.5	236.0	
7	$D+P$	C	55.9	236.8	64.7
8	$D+C$	C	54.4	237.0	55.4
9	$D+S$	C	59.2	236.9	66.6
10	$D+E$	C	25.4	237.5	37.3
11	$D+C+P$	C	50.7	236.9	52.6
12	$D+S+P$	C	55.5	236.9	63.7
13	$D+E+P$	C	21.8	237.4	33.3
14	$D+C+P+E$	C	6.7	238.6	14.5
15	$D+S+P+E$	C	19.9	237.5	31.1
16	$D+S+P+C$	C	49.8	237.0	50.5
17	$D+P+C+S+E$	C	4.3	239.0	11.5
18	$D+P+C+S+E+C:E$	C	2.8	250.8	12.9
19	$D+P+C+S+E+S:E$	C	6.2	239.1	13.4
20	$D+P+C+S+E+S:C$	C	1.7	239.0	14.9
21	$D+P+C+S+E+S:C+C:E$	C	0.0	252.0	16.1
22	$D+P+C+S+E+S:C+S:E$	C	2.7	240.7	17.8
23	$D+P+C+S+E+E:C+S:E$	C	4.7	251.2	14.8
24	$D+P+C+S+E+E:C+S:E+C:S$	C	0.7	271.8	18.2

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

MRDS - Exercise

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!

MRDS - Summary

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

Survey Protocol

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

MRDS - Summary

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.

MRDS - Summary

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.