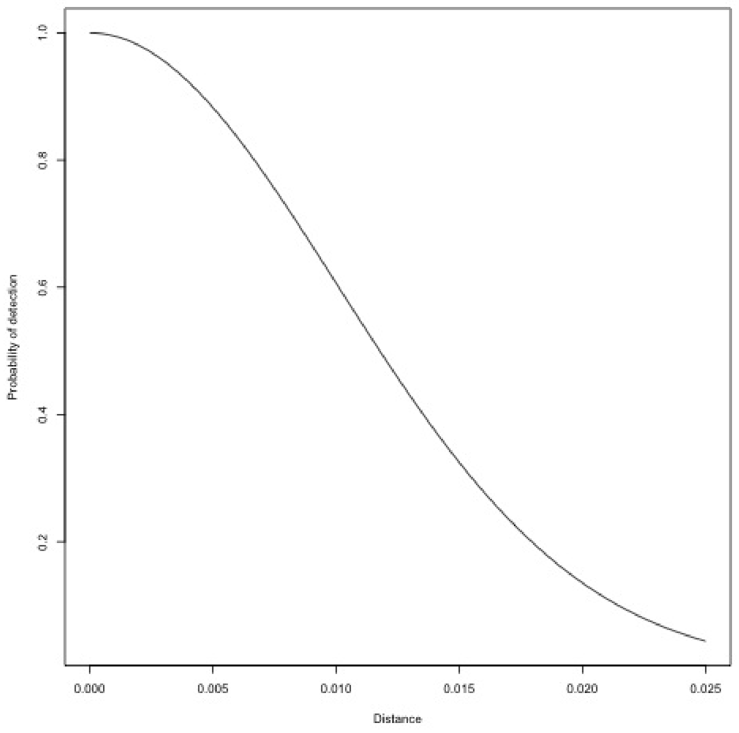
Distance sampling: Advanced topics

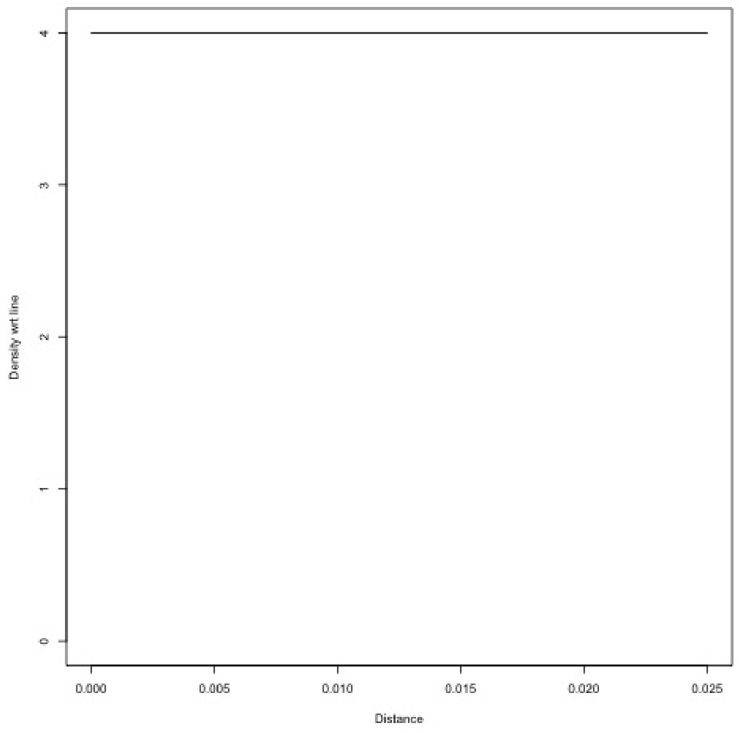
David L Miller

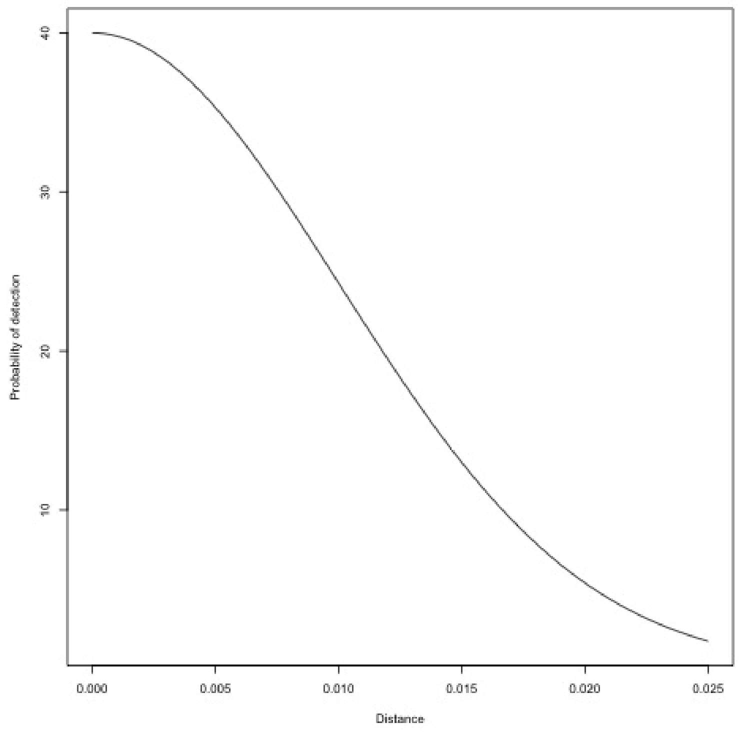
Recap

Line transects - general idea

- Calculate *average detection probability*
  - using detection function ($g(x)$)
- $\hat{p} = \int_0^w \frac{1}{w} g(x; \hat{\theta}) dx$
- $\frac{1}{w}$ tells us about assumed density wrt line
  - *uniform* from the line (out to w)







Line transects - distances

- Model drop-off using a *detection function*
- Use extra information estimate $\hat{N}$
- How should we adjust n? (inflate by $n/\hat{p}$))

Fitting detection functions

- Using the package `Distance`
- Need to have data setup a certain way
  - At least columns called `object`, `distance`

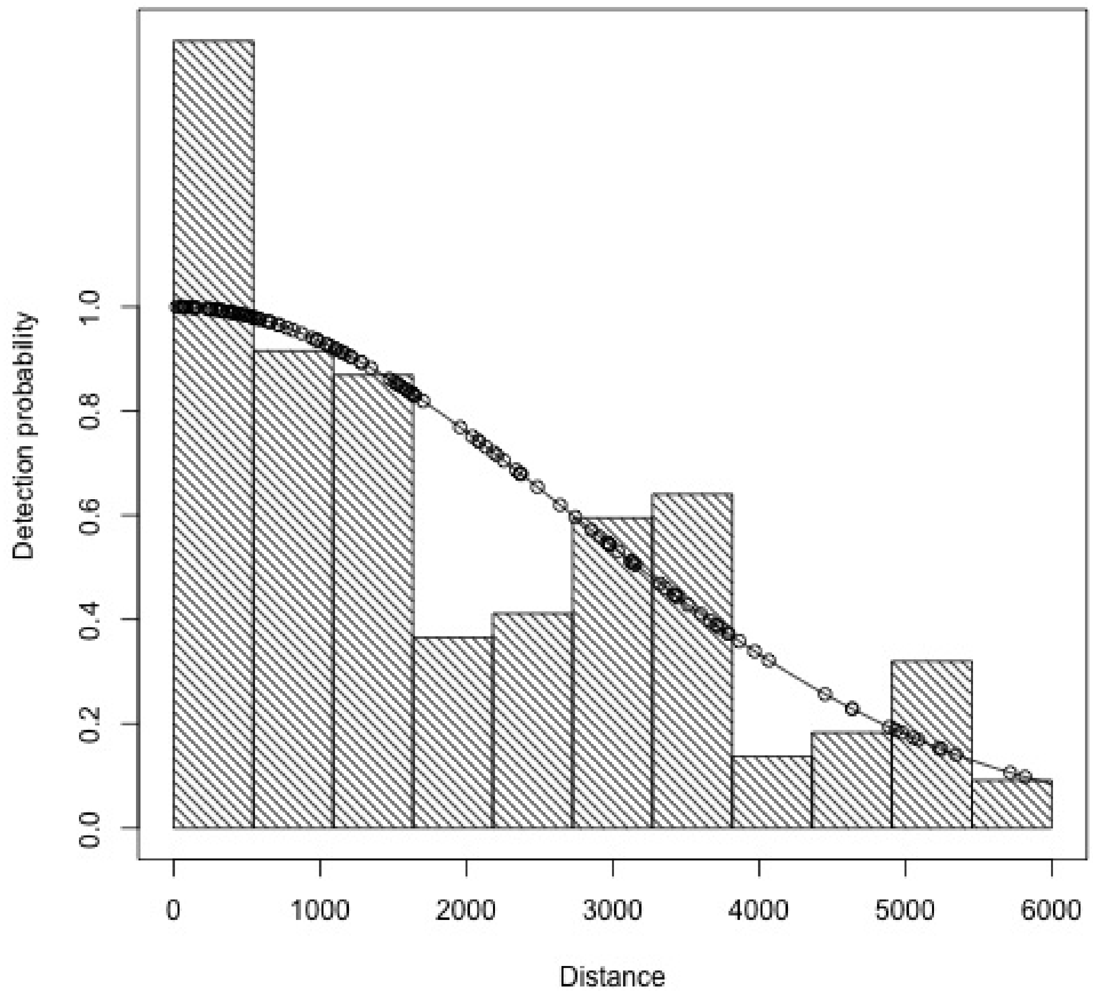
```
library(Distance)
df_hn <- ds(distdata, truncation=6000, adjustment = NULL)
```

Model summary

```
summary(df_hn)
```

```

Summary for distance analysis
Number of observations :  132
Distance range         :  0  -  6000

Model : Half-normal key function
AIC   : 2252.06

Detection function parameters
Scale Coefficients:
            estimate         se
(Intercept) 7.900732 0.07884776

                        Estimate          SE         CV
Average p              0.5490484  0.03662569 0.06670757
N in covered region  240.4159539 21.32287580 0.08869160
```

Plotting models




```
plot(df_hn)
```

New stuff

Overview

Here we'll look at:

- Model checking and selection
- What else affects detection?
- Estimating abundance and uncertainty
- More R!

Why check models?

- AIC best model can still be a terrible model
- AIC only measures **relative** fit
- Don't know if the model gives “sensible” answers

What to check?

- Convergence
  - Fitting ended, but our model is not good
- Monotonicity
  - Our model is "lumpy"
- "Goodness of fit"
  - Our model sucks statistically
- (Other sampling assumptions are also important!)

Convergence

`Distance` will warn you about this:

```
** Warning: Problems with fitting model. Did not converge**
Error in detfct.fit.opt(ddfobj, optim.options, bounds,
misc.options) :
  No convergence.
```

This can be complicated, see `?"mrds-opt"` for info.

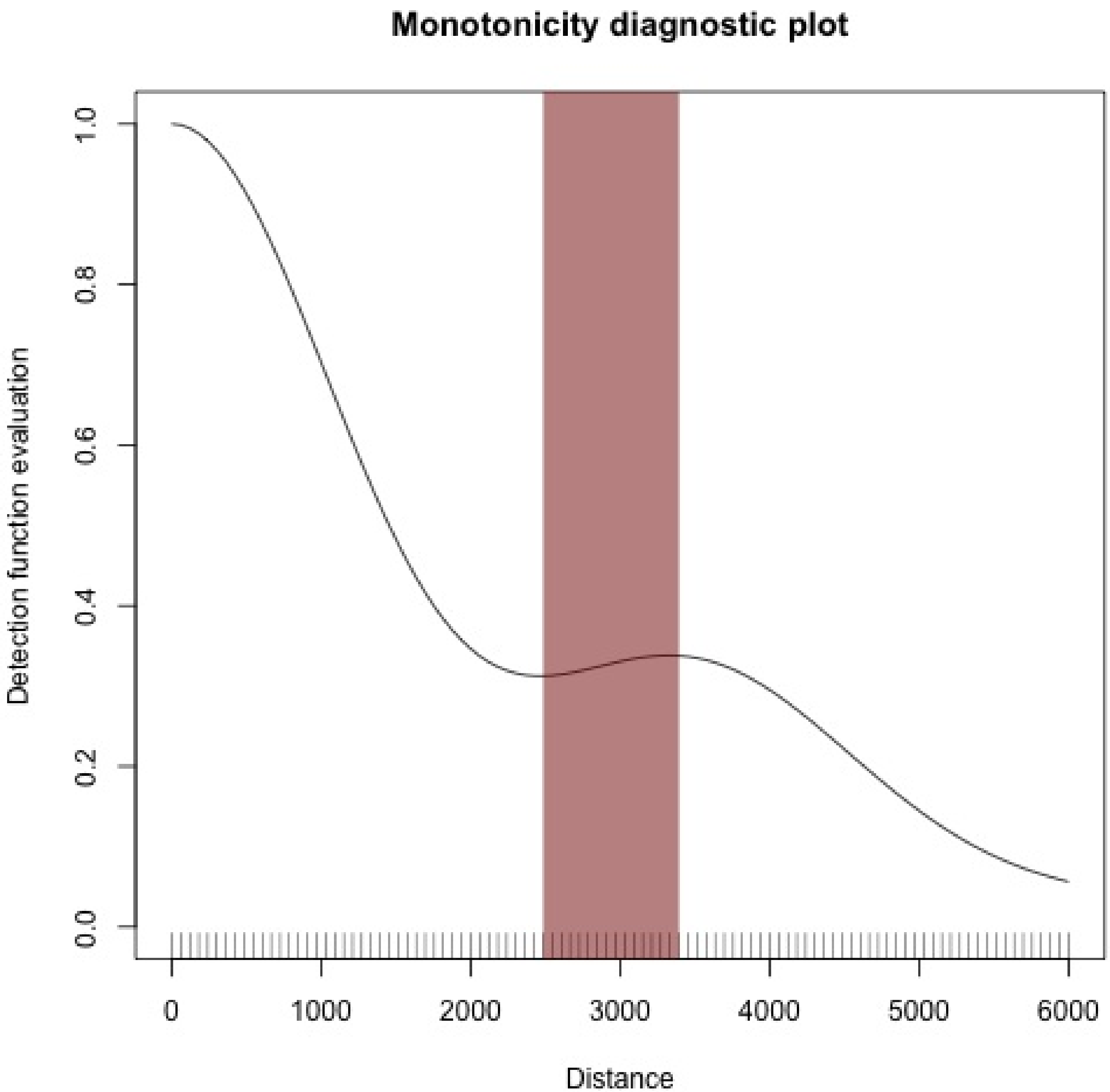
Monotonicity

- Only a problem with adjustments
- `check.mono` can help

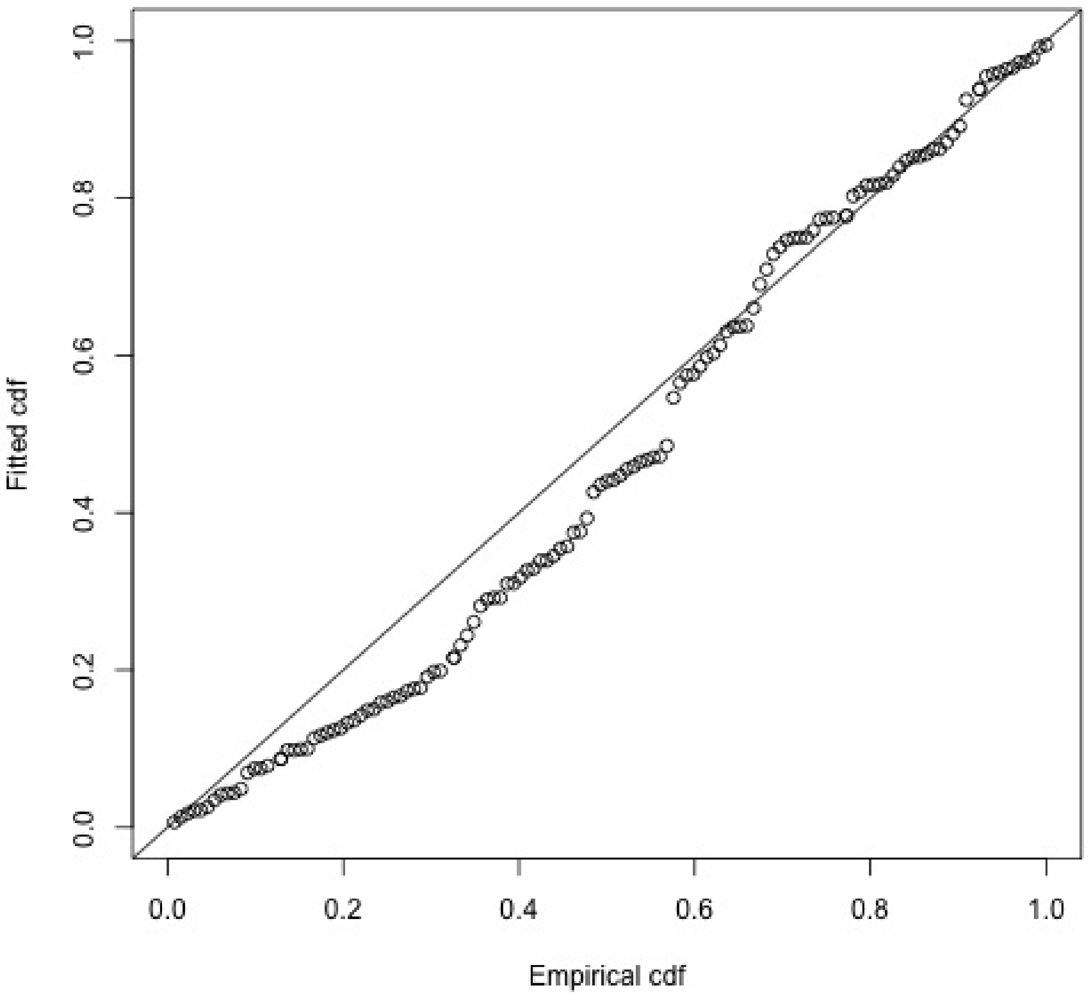
```
check.mono(df_hr$ddf)
```

```
[1] TRUE
```

Monotonicity (when it goes wrong)



Goodness of fit



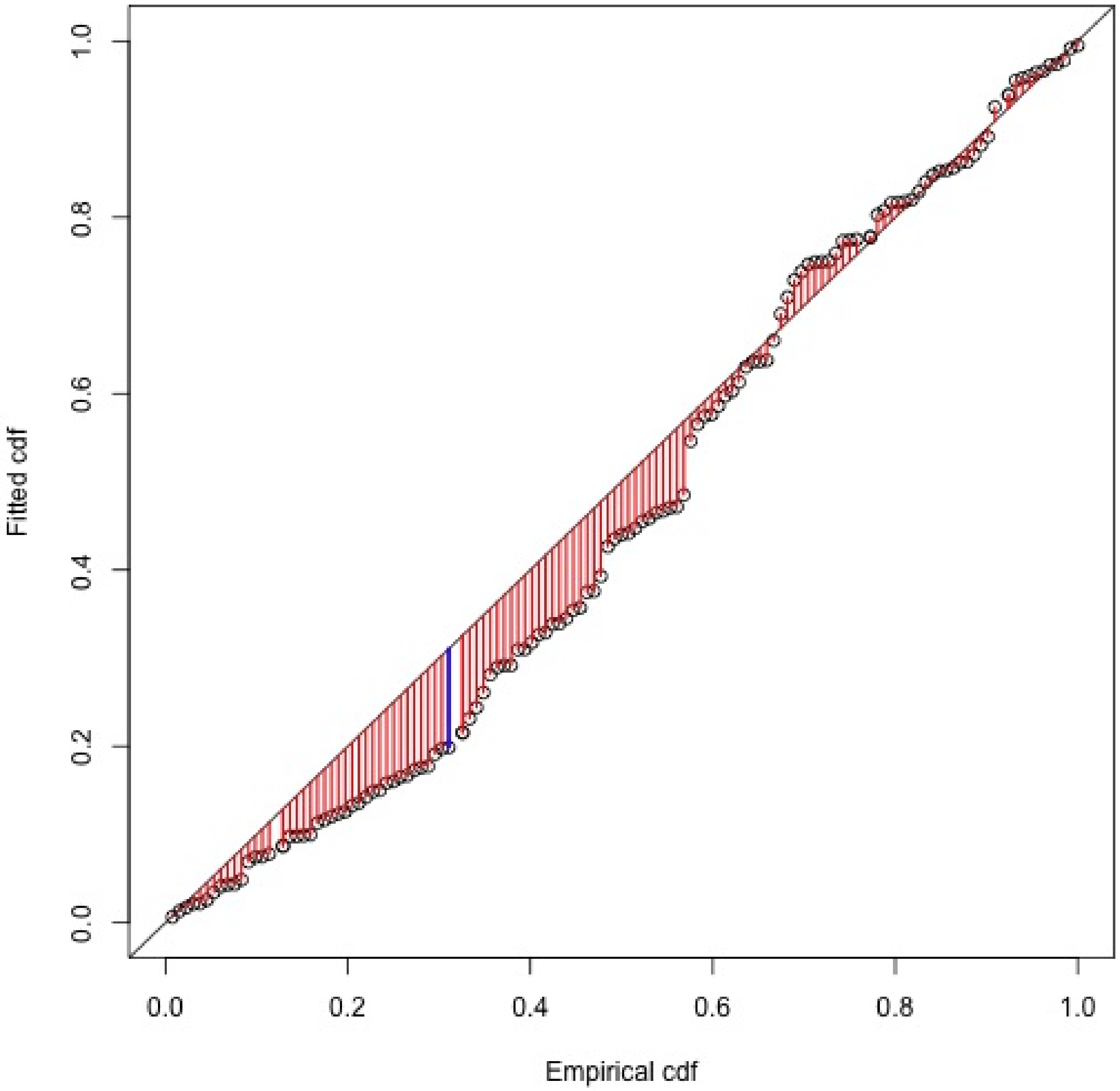

```
ddf.gof(df_hn$ddf)
```

- Check fitted distribution of distances matches empirical
- # distances below distance vs. # observations below given cumulative probability

Goodness of fit

- As well as quantile-quantile plot, tests
- Absolute measure of fit (vs. AIC)
- Kolmogorov-Smirnov: largest distance on Q-Q plot
- Cramer-von Mises: tests sum of distances

Goodness of fit



- blue: Kolmogorov-Smirnov
- red: Cramer-von Mises

Detection function model selection

- Fit models
- Look at `summary` and `plot` (fitting issues?)
- Look at goodness of fit results, `ddf.gof`
- AIC to select between models
  - Parsimonous: “robust” and “efficient” models

Example: fitting detection functions

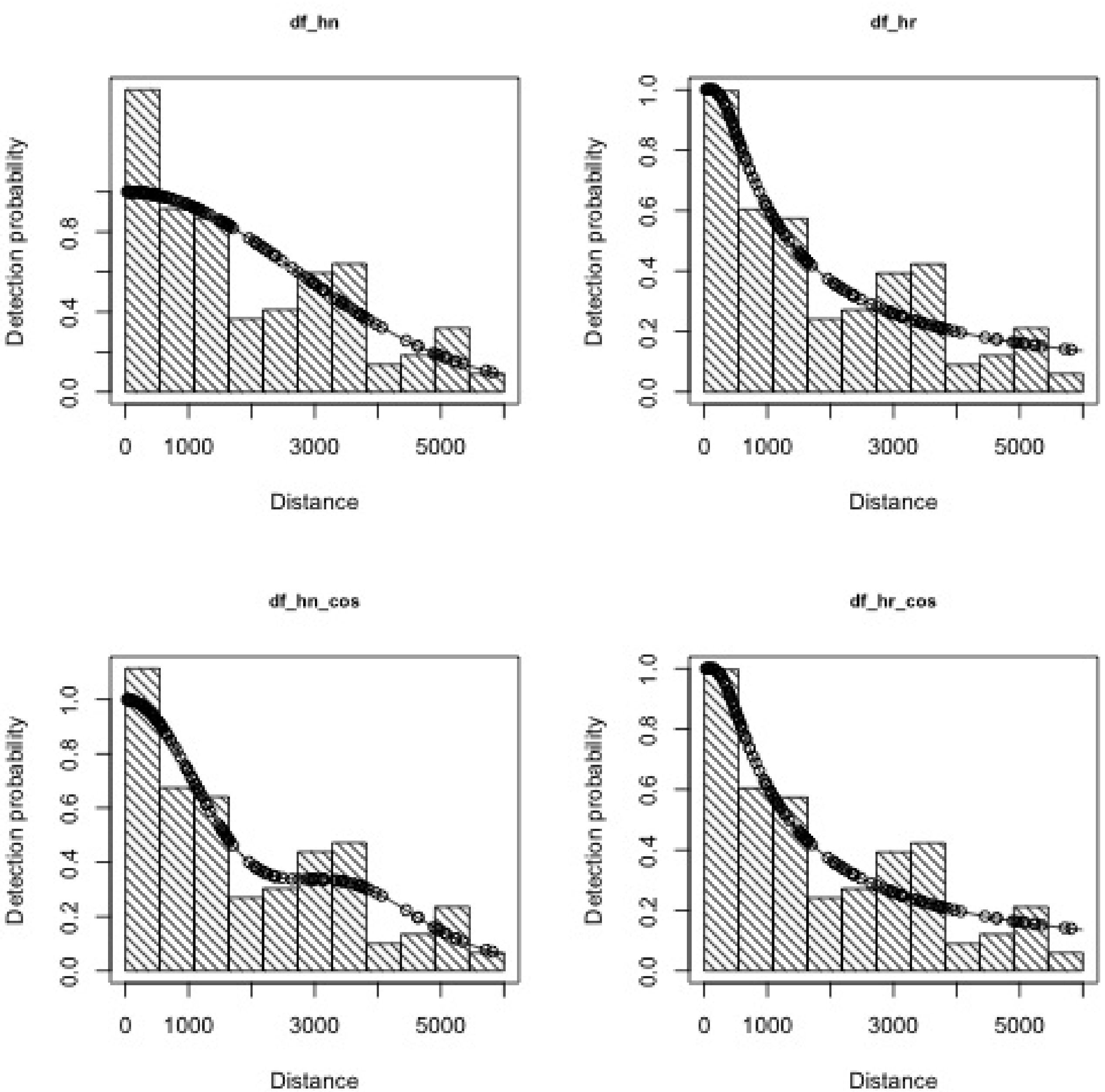
```
df_hn <- ds(distdata, truncation=6000, adjustment = NULL)
```

```
df_hn_cos <- ds(distdata, truncation=6000, adjustment = "cos")
```

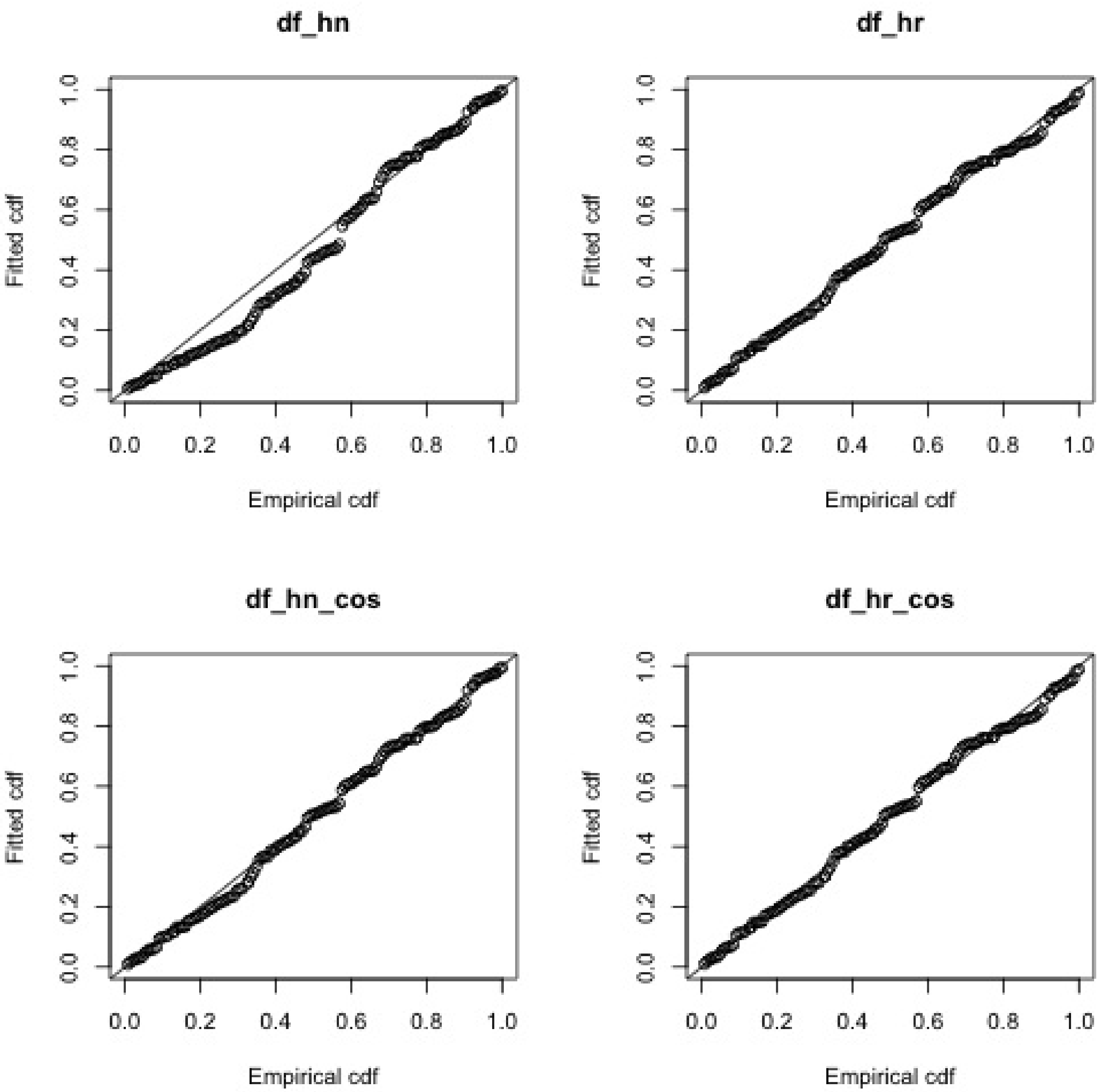
```
df_hr <- ds(distdata, truncation=6000, key="hr", adjustment =
NULL)
```

```
df_hr_cos <- ds(distdata, key="hr",  truncation=6000, adjustment =
"cos")
```

Plotting those models



Q-Q plots


df_hn
df_hr
df_hn_cos
df_hr_cos
Fitted cdf
Empirical cdf
0.0
0.2
0.4
0.6
0.8
1.0

AIC

```r
df_hn$ddf$criterion
```

```
[1] 2252.06
```

```r
df_hn_cos$ddf$criterion
```

```
[1] 2247.69
```

```r
## same model!
df_hr$ddf$criterion
```

```
[1] 2247.594
```

```r
df_hr_cos$ddf$criterion
```

```
[1] 2247.594
```

Selection

- Not much between these models!
- You'll get to investigate these and more in the lab

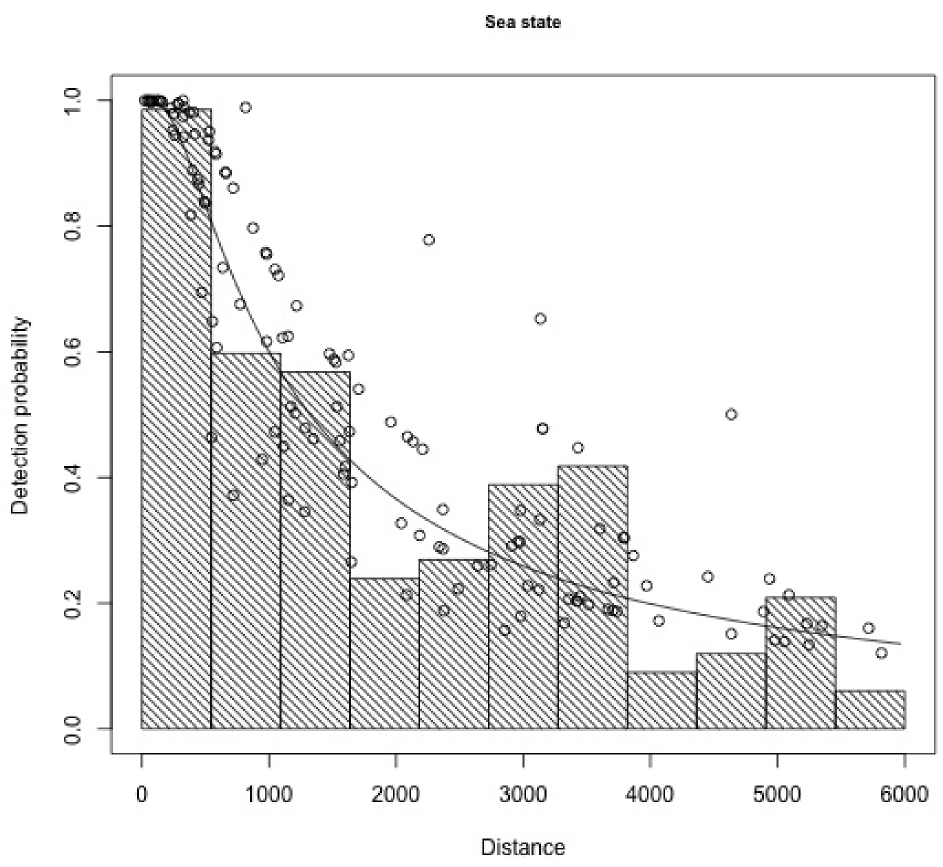
What else affects detectability?

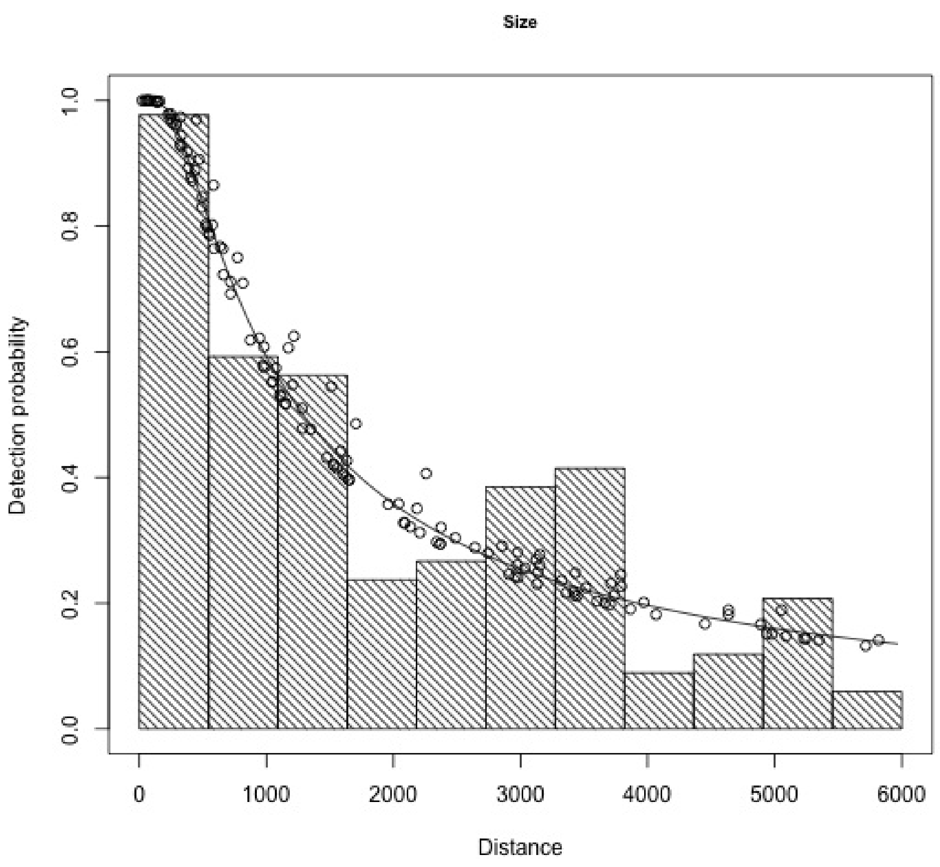
Covariates

- Observer characteristics
  - observer name
  - platform
- Animal characteristics
  - sex
  - size
  - group size
- Weather conditions
  - sea state
  - glare
  - fog

How do we include covariates?

- Affects scale, not shape





Covariates in the scale

$$\exp\left(\frac{-x^2}{2\sigma^2}\right) \text{ or } 1-\exp\left[\left(\frac{-x}{\sigma}\right)^{-b}\right]$$

Decompose $\sigma = \exp\left(\beta_0 + \beta_1 z_1 + \ldots\right)$

What does detectability mean?

- $\hat{p}$ is now $\hat{p}_i$ (or $\hat{p}(\mathbf{z}_i)$)
- Average probability of detection (average over *distances*)
- Also calculate an average $\hat{p}$ as a summary

Covariates in R

- Add `formula=...` to our `ds()` call:

```
df_hr_ss <- ds(distdata, truncation=6000,
               key="hr", formula=~SeaState)
```

```
df_hr_ss_size <- ds(distdata, truncation=6000,
                    key="hr", formula=~SeaState+size)
```

Summaries of covariate models

```r
summary(df_hr_ss)
```

```

Summary for distance analysis
Number of observations :  132
Distance range         :  0  -  6000

Model : Hazard-rate key function
AIC   : 2247.347

Detection function parameters
Scale Coefficients:
              estimate        se
(Intercept)  8.1019226 0.7906353
SeaState    -0.4473291 0.2797965

Shape parameters:
              estimate        se
(Intercept) 0.07319982 0.2417426

                        Estimate          SE        CV
Average p              0.3583687  0.07308615 0.2039412
N in covered region 368.3357858 79.54571167 0.2159598
```

”Average p”

$$p(\mathbf{z}_i) = \int_0^w g(x; \boldsymbol{\theta}, \mathbf{z}_i)dx \quad \text{for } i = 1, \ldots, n$$

```
unique(predict(df_hr_ss$ddf)$fitted)
```

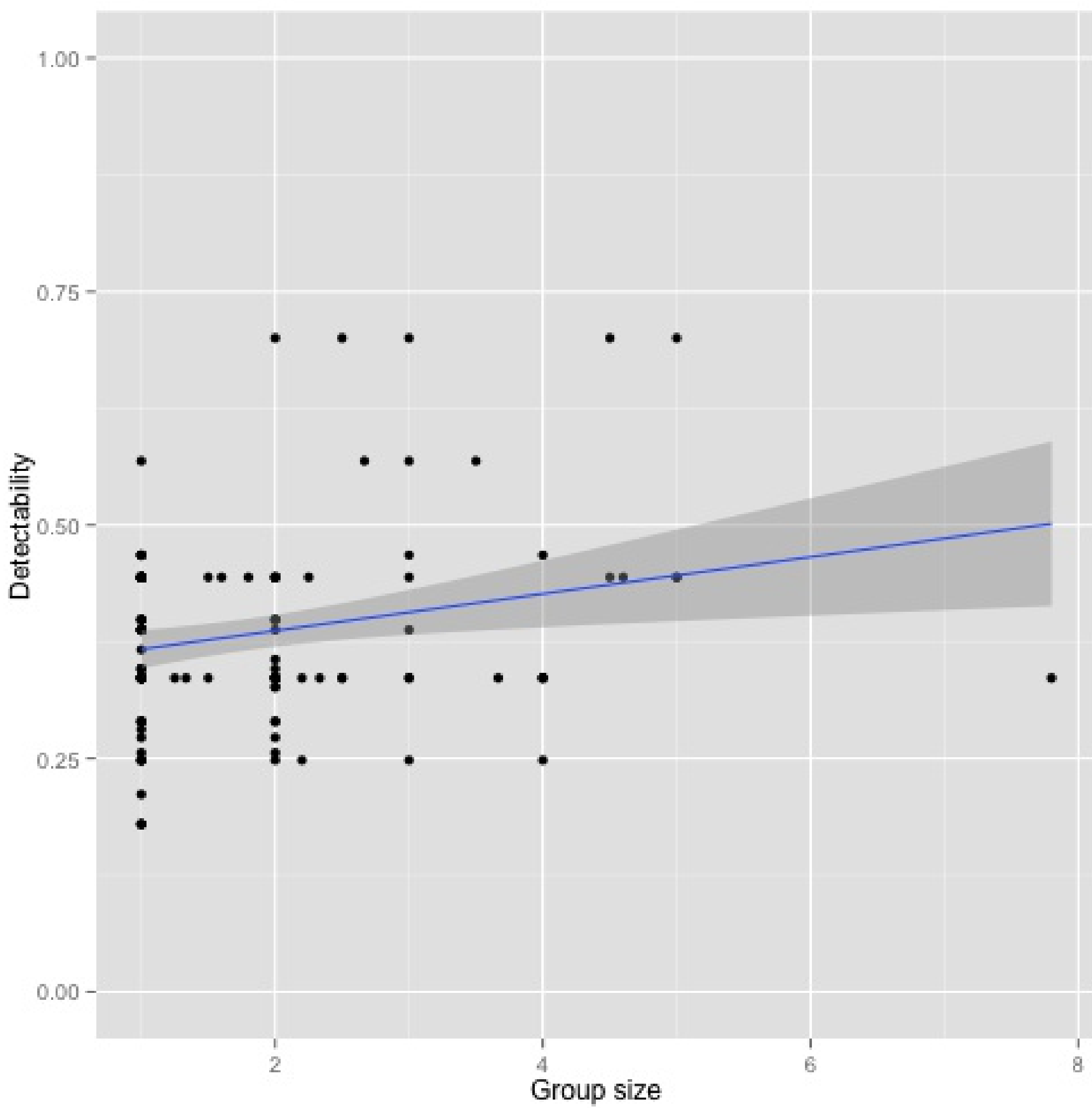
```
 [1] 0.3360342 0.3876026 0.2895189 0.2480620 0.3985064 0.4439768
0.2723358
 [8] 0.2559550 0.2808264 0.3459473 0.3263237 0.3663789 0.5684780
0.2114896
[15] 0.3560627 0.4677557 0.1795108 0.7000862
```

Group size

What are groups?

- *Functional* definition (NO ecology!)
  - If animals are near each other, they are in a group
- This probably affects detectability
  - Bigger groups $\Longrightarrow$ easier to detect
- Two inferential targets
  - abundance of groups
  - abundance of individuals

Detection and group size



- Not a huge change here
- Bigger effect for animals that occur in large groups
  - Seabirds
  - Dolphins

Estimating abundance

Estimating abundance

- As before, assume density same in sampled/unsampled area
- Horvitz-Thompson estimator

$$\hat{N} = \frac{A}{a} \sum_{i=1}^{n} \frac{s_i}{\hat{p}_i}$$

where s_i is group size, n is number of *observations* (groups)

Estimating uncertainty

Sources of uncertainty

$$\hat{N} = \frac{A}{a} \sum_{i=1}^{n} \frac{s_i}{\hat{p}_i}$$

- Uncertainty in n is from sampling
- Uncertainty in $\hat{p}$ is from the model

Uncertainty from sampling

- Usually calculate *encounter rate* variance
- Encounter rate is n/L
- (Measure of spatial variability $\Longrightarrow$ uncertainty)
- "Objects per unit length of transect surveyed"
- Fewster et al. (2009) is the definitive reference

Uncertainty from the model

- Model uncertainty from estimating parameters
- Maximum likelihood theory gives uncertainty in model pars

Putting those parts together

Obtain overall CV by adding squared CVs:

$$\mathrm{CV}^2\left(\hat{\mathrm{D}}\right) \approx \mathrm{CV}^2\left(\frac{\mathrm{n}}{\mathrm{L}}\right) + \mathrm{CV}^2\left(\hat{\mathrm{p}}\right)$$

(Running through this quickly, see bibliography for more details)

(One other thing...)

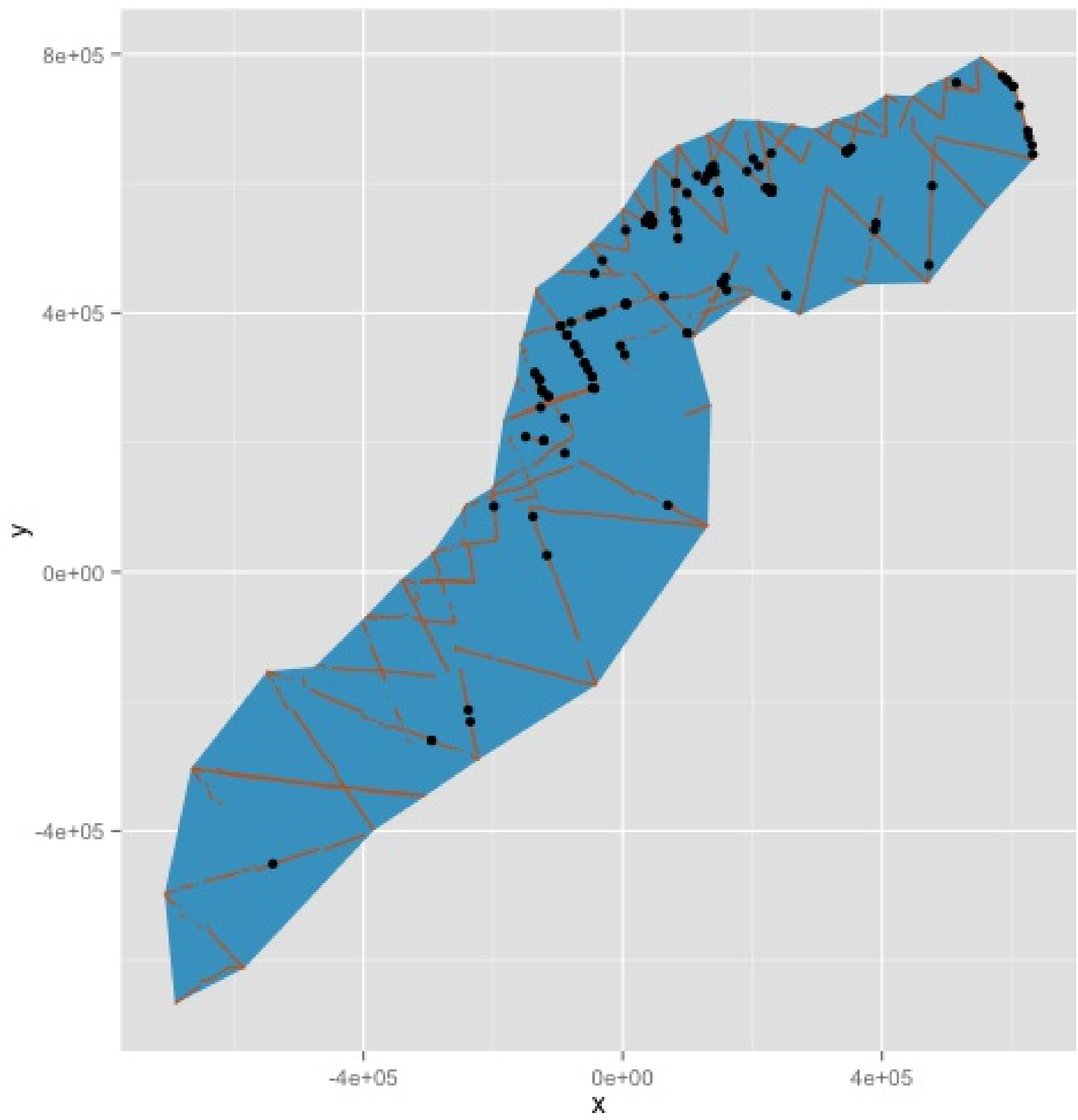
- Assume that group size is recorded correctly
- This is almost never true
- There are ways to deal with this
- See bibliography for more details

Variance and abundance in R…

Data required

- Need three tables
  - region: whole area
  - sample: the samples (transects)
  - observation: relate samples to observations

Schematic



- region
- sample
- observations

Region table

```
head(region.table)
```

```
  Region.Label       Area
1    StudyArea 5.285e+11
```

Sample table

```r
head(sample.table)
```

```
      Sample.Label    Effort Region.Label
1 en0439520040624 144044.67    StudyArea
2 en0439520040625 167646.84    StudyArea
3 en0439520040626  59997.33    StudyArea
4 en0439520040627  33821.89    StudyArea
5 en0439520040628 147414.92    StudyArea
6 en0439520040629 101107.83    StudyArea
```

Observation table

```
head(obs.table)
```

```
  object    Sample.Label Region.Label
1      1 en0439520040628    StudyArea
2      2 en0439520040628    StudyArea
3      3 en0439520040628    StudyArea
4      4 en0439520040628    StudyArea
5      5 en0439520040629    StudyArea
6      6 en0439520040629    StudyArea
```

Abundance and variance

This generates a lot of output (here is a snippit):

```
dht(df_hr$ddf, region.table, sample.table, obs.table)
```

```
Summary for individuals

Summary statistics:
     Region      Area  CoveredArea  Effort     n           ER
se.ER     cv.ER mean.size
1 StudyArea 5.285e+11 113981689066 9498474 238.7 2.513035e-05
5.667492e-06 0.2255238  1.808333
    se.mean
1 0.1020928

Abundance:
  Label Estimate       se        cv      lcl      ucl       df
1 Total 3053.558 943.7425 0.3090632 1682.187 5542.912 170.9157
```

More investigation in the practical exercises…

From that summary...

- Individuals observed: $n = 238.7$
- Covered area: $a = 113,981,689,066 m^2$
- Study area: $A = 5.285 \times 10^{11} m^2$
- Detectability: $\hat{p} = 0.3625$

So

$$\hat{N} = \frac{n}{\hat{p}} \frac{A}{a} = 3053.558$$

Recap

Summary

- How to check detection function models
- Covariates can affect detectability
- Group size
- Sources of uncertainty
- Estimation of abundance and variance