Tweedie versus $\log(y+\epsilon)$ for modelling zero-inflated continuous data

Russell Millar

University of Auckland

Background

The Tweedie family of distributions is able to handle zero-inflated continuous data. They date back to 1984, and my use of them to 2020 for analysis of a complex M-BACI design to assess effects of a marine seismic survey on fish catch rate.

Background

I am now using the Tweedie for comparison of catch rates when trawl gear is altered. Others continue to use $\log(y + \epsilon)$ approach (and in many other areas where zero-inflated continuous data are encountered) .

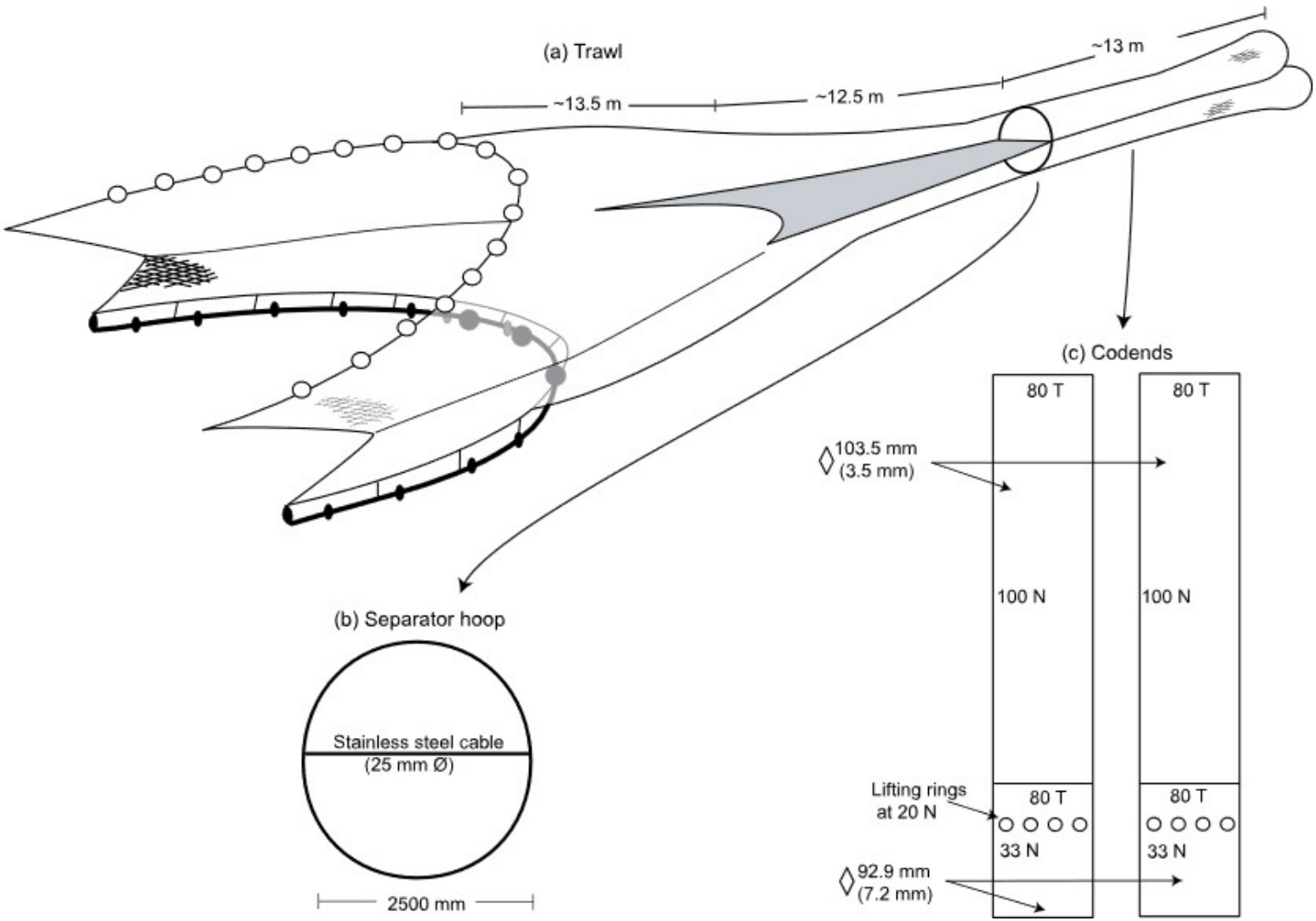
Suitable for a graduate research project with a U. Auckland data science student.

Became an international research project[1] subsequent to joining a delegation to India in Feb this year to promote joint graduate programs between U. Auckland and Indian universities at ITs.

[1] With Professor Tathagata Bandyopadhyay and students Siddarth and Naisarg at IICT Gandhinagar, Gujarat, India.

Motivating Example

Weight of fish caught in upper or lower part of a split-codend trawl.



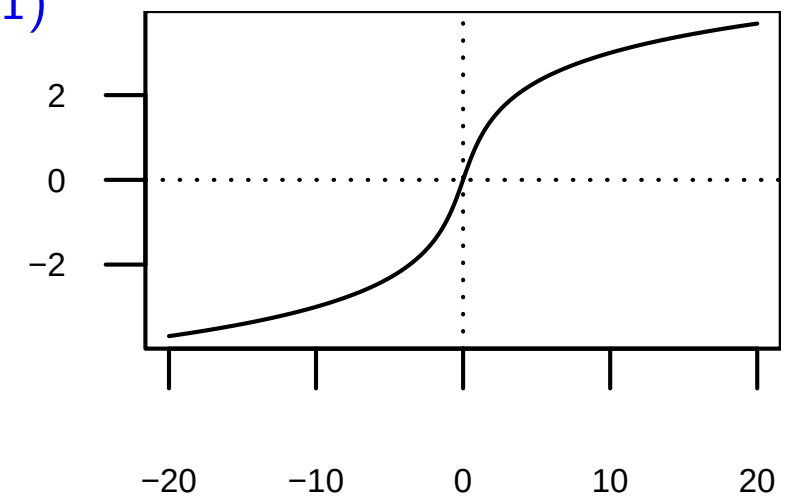
Motivating Example

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	Set.Id	Start.Date	Experimen	Treatment	Start.Time	Total.catc	Eastern.ar	Gould.squi	John.dory.	Morwong.	Redfish.(c)	Red.gurna	Silver.trev	Tiger.flath	Eastern.sc	Longfin.bigeye.(c).wt	
2	7775	28/04/2022	One	Lower	5:00:36	206.1	20	0.6	0	0.1	0	0.3	0	42	2	0	
3	7776	28/04/2022	One	Lower	8:22:43	225.1	20	0	0	0	0	6	1.5	20.5	6	0	
4	7777	28/04/2022	One	Lower	11:45:51	374.7	136	0	0	0	0	3.7	0	44	1.8	0	
5	7778	29/04/2022	One	Lower	5:01:53	188	18	2	0.9	0	0	23	0	34.8	0	0	
6	7779	29/04/2022	One	Lower	8:40:21	192.4	5	2	2	0.2	0	15	7	40	0	0	
7	7780	29/04/2022	One	Lower	11:45:40	328.2	45	0.3	1.8	3	5.5	7	0.4	51	0	0	
8	7781	29/04/2022	One	Lower	14:45:28	257.7	0	0	2	5.6	0	6.5	0	19	0	0	
9	7782	1/05/2022	One	Lower	4:47:32	1672.8	0	42	0	0.6	0	0	0	80	0	0	
10	7783	1/05/2022	One	Lower	8:52:42	927.9	0	5	0	0.6	0	0	0	0	0	0	
11	7784	1/05/2022	One	Lower	13:45:13	118	0	1.6	1.7	8.8	0.2	0.6	0	27.9	6.5	0	
12	7785	2/05/2022	One	Lower	4:45:23	683.1	0	17	0	0.6	0.7	0	0	92	0	0	
13	7786	2/05/2022	One	Lower	9:10:00	245.9	0	1.8	2.9	0.4	0	6.4	43	4.9	0	1.1	
14	7787	2/05/2022	One	Lower	11:44:56	364.4	0	2.3	2.5	7.5	0.5	2.5	40	18.7	0.2	0	
15	7788	2/05/2022	One	Lower	14:29:42	237.7	1	0	4.5	3.6	0	2.9	0.8	2.3	3.5	0.6	
16	7789	3/05/2022	One	Lower	4:35:33	287.3	0	5.1	0	1.1	0	0.9	0	33.1	0	0	
17	7790	3/05/2022	One	Lower	8:25:30	425.3	0	0	0	4.2	0.8	0	0	30	0	0	
18	7791	3/05/2022	One	Lower	13:06:47	263.1	54.5	0.9	0.2	0.4	0	16	0	39.7	0.1	0	
19	7775	28/04/2022	One	Upper	5:00:36	52.3	5	4	0	0	0	0	0	9.2	0.6	0	
20	7776	28/04/2022	One	Upper	8:22:43	60.5	8	0	0	0	0	2.5	1.8	2.8	0.5	0	
21	7777	28/04/2022	One	Upper	11:45:51	46.5	9	1	1	0	0	2	1.8	3	0.3	0	
22	7778	29/04/2022	One	Upper	5:01:53	84.8	0	2.2	0.8	0	0.2	17	0	12	0	0	
23	7779	29/04/2022	One	Upper	8:40:21	86.4	9.5	0.8	1.5	0	0	15.5	6	29.2	0	0	
24	7780	29/04/2022	One	Upper	11:45:40	84.9	7	0	0	0.2	0.9	0	0.5	18.3	0	0	
25	7781	29/04/2022	One	Upper	14:45:28	137	0	0	0.7	0	0	4.5	0	17.2	0	0	
26	7782	1/05/2022	One	Upper	4:47:32	262.5	0	54	1	0	0	0	0	15	0	0	
27	7783	1/05/2022	One	Upper	8:52:42	300.1	0	12	0	0	0	0	0	1	0	0	
28	7784	1/05/2022	One	Upper	13:45:13	37.5	0	2.2	0.5	0	0	3.5	0	6	0.5	6.5	
29	7785	2/05/2022	One	Upper	4:45:23	412.4	0	15	0.6	0	0	3.2	0	18	0	0	
30	7786	2/05/2022	One	Upper	9:10:00	109.15	0	8	1.7	1.3	0	12.4	31	4.5	0.05	1.2	
31	7787	2/05/2022	One	Upper	11:44:56	99.6	0	4	0.3	0.5	0.2	5.9	14	9	1.3	0.7	
32	7788	2/05/2022	One	Upper	14:29:42	57.4	0	3.5	1.3	0.1	0	5.8	0.5	1	27	1.1	
33	7789	3/05/2022	One	Upper	4:35:33	75.3	0	1.8	0	0	0	0.7	0	2.6	0	0	
34	7790	3/05/2022	One	Upper	8:25:30	178.4	0	2	0	0	0	0	0	1.5	0	0	
35	7791	3/05/2022	One	Upper	13:06:47	75.6	0	2.9	0.8	0	0	8.5	0.4	28.7	0	1.7	

Analysis options

Assume that the data contain at least one zero and let m^+ be the minimum positive value. Possible approaches:

- Zero-inflated log-normal.
- Model $\log(y_i + \epsilon)$ using a linear model. Choice of ϵ?
  - $\log(y_i + 1)$
  - $\log(y_i + \widehat{\epsilon})$ with $\widehat{\epsilon}$ found by MLE
  - $\log(y_i + m^+)$
  - $\log(y_i + \frac{m^+}{2})$
  - $\log(y_i + \sqrt{y_i^2 + 1})$



- Tweedie model

Zero-inflated log-normal

- Model the proportion of zeros using a logit model, say.
- For $y > 0$, model $\log(y)$ using a linear model.

In practice the fishery technologist wants simple answers to questions such as:

> *What is the difference in (expected) catch between the upper and lower codend?*
>
> *Does the difference in (expected) catch between the upper and lower codend depend of the presence of a horizontal separator panel?*

Model $\log(y + \epsilon)$ using a linear model.

- $\log(y_i + 1)$. **Not scale invariant.**
- $\log(y_i + \widehat{\epsilon})$ **Too complex.** **Tail wagging the dog.**
- $\log(y_i + m^+)$
- $\log(y_i + \frac{m^+}{2})$
- $\log(y_i + \sqrt{y_i^2 + 1})$ **Not scale invariant.**

Lack of scale invariance means that recording catch in kilograms will give a different result than if catch was recorded in tonnes or grams.

Tweedie distribution

The Tweedie[2] is an exponential dispersion family model of the form:

$$f(y;\mu,\phi,\rho) = a(y,\phi)\exp\left(\frac{y\theta-\kappa(\rho,\theta)}{\phi}\right)$$

where $\mu = \kappa'(\rho,\theta)$. This density function has no closed form.

In the `tweedie` R package it is evaluated by approximation to an infinite series expansion, or using Fourier inversion.

The fitting of Tweedie models is now implemented in several R packages, including `glmmTMB` and `cplm`.

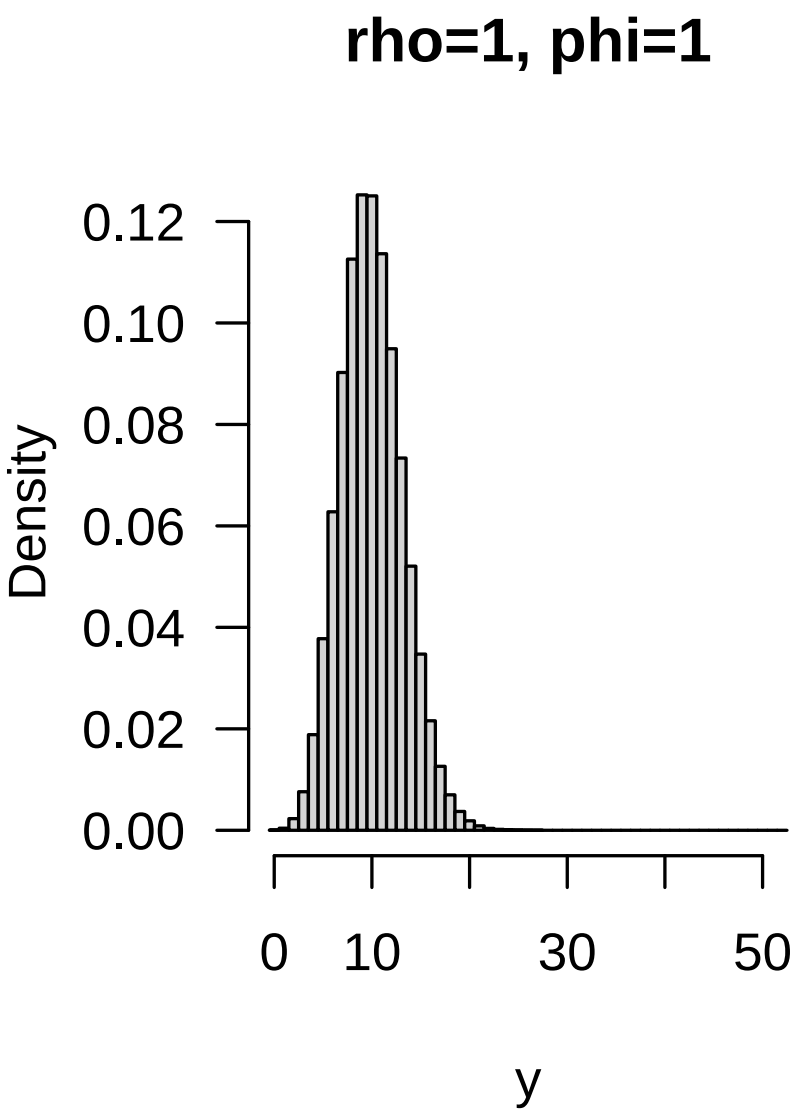
[2] Named after Maurice Tweedie's 1984 paper in ISI conference proceedings.

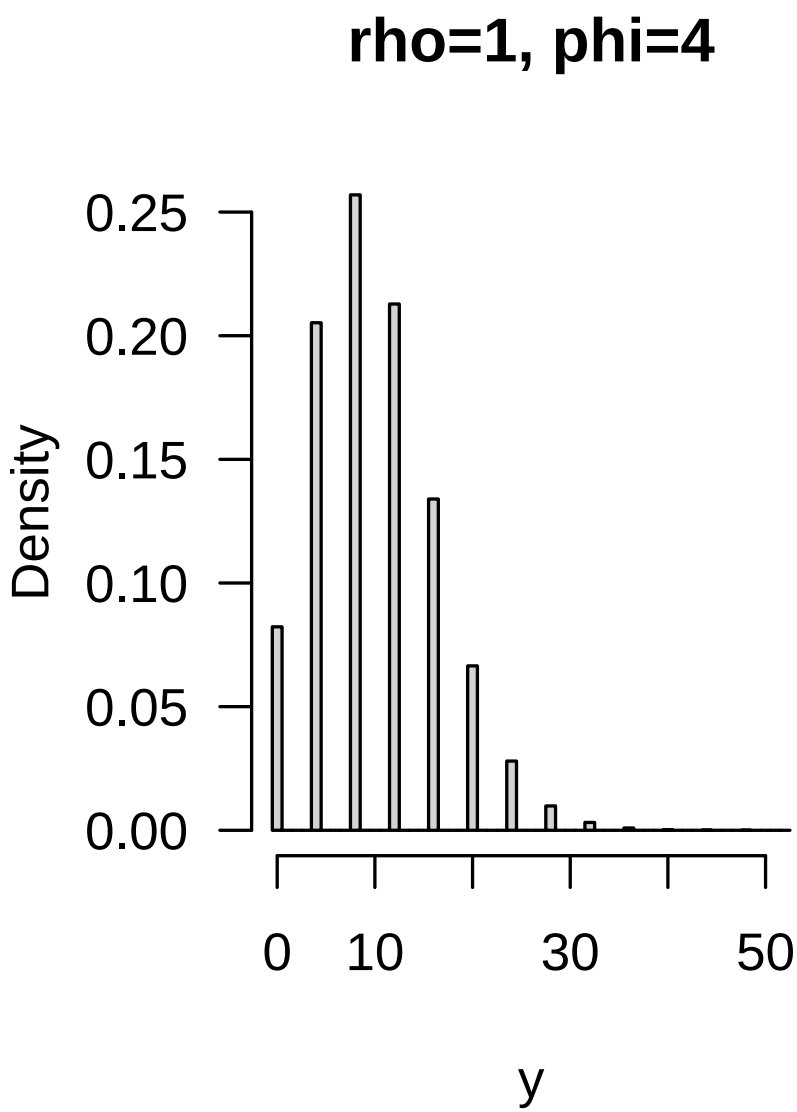
Tweedie properties

Some properties:

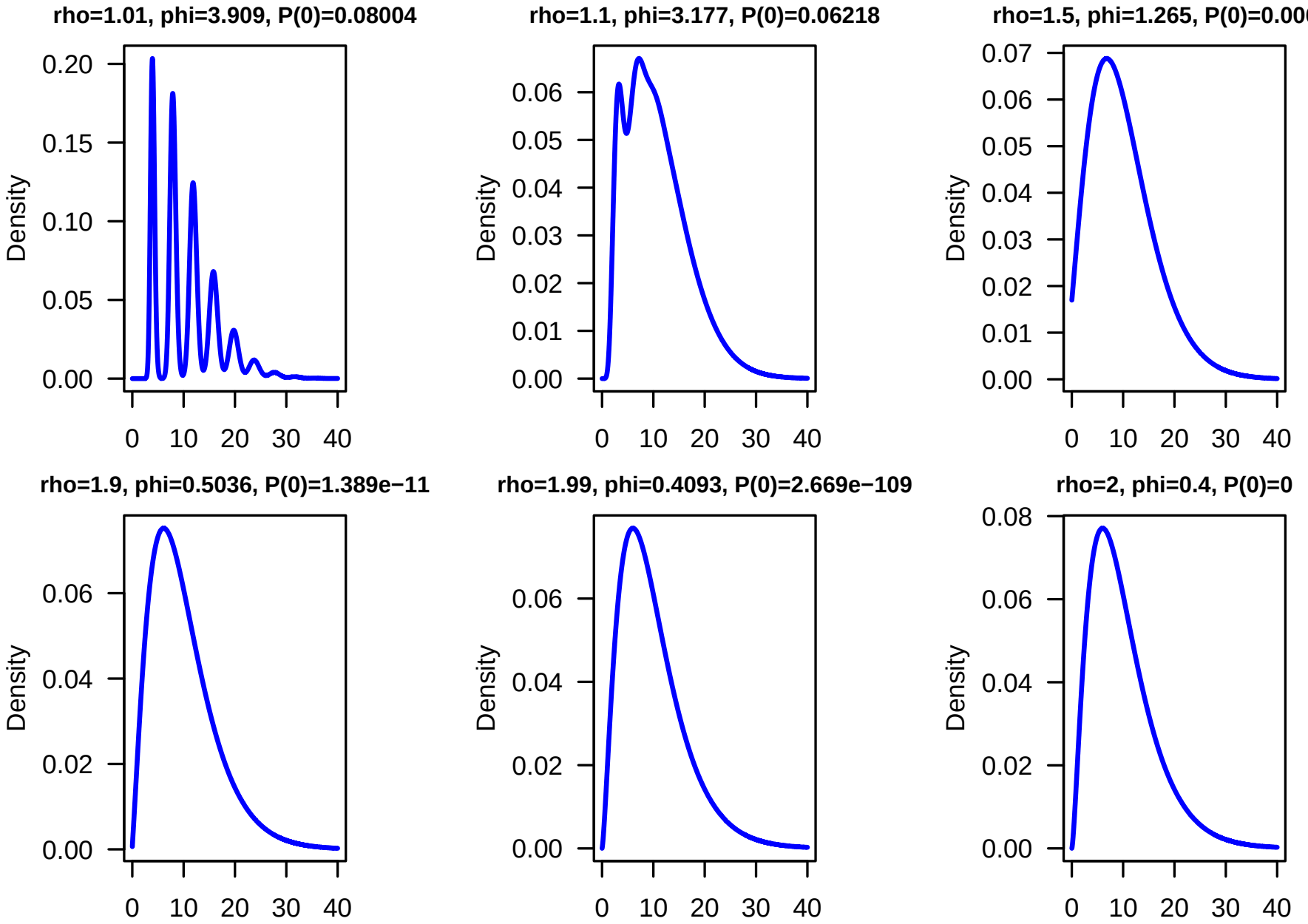
- Three parameters, μ (mean), ϕ (dispersion), ρ (power)
- Satisfies Taylor's power law: $\text{var}(Y) = \phi\mu^{\rho}$
- $\text{Prob}(Y = 0) = \exp(-\mu^{(2-\rho)}/(\phi(2-\rho))), \quad 0 < \rho < 1$
- $\rho = 0$: normal distribution
- $\rho = 1, \phi = 1$: Poisson
- $\rho = 1, \phi \neq 1$: quasi-Poisson
- $1 < \rho < 2$: Poisson mixture of gamma distributions
- $\rho = 2$: gamma distribution
- $\rho = 3$: inverse Gaussian distribution

Tweedie examples: $\mu = 10$

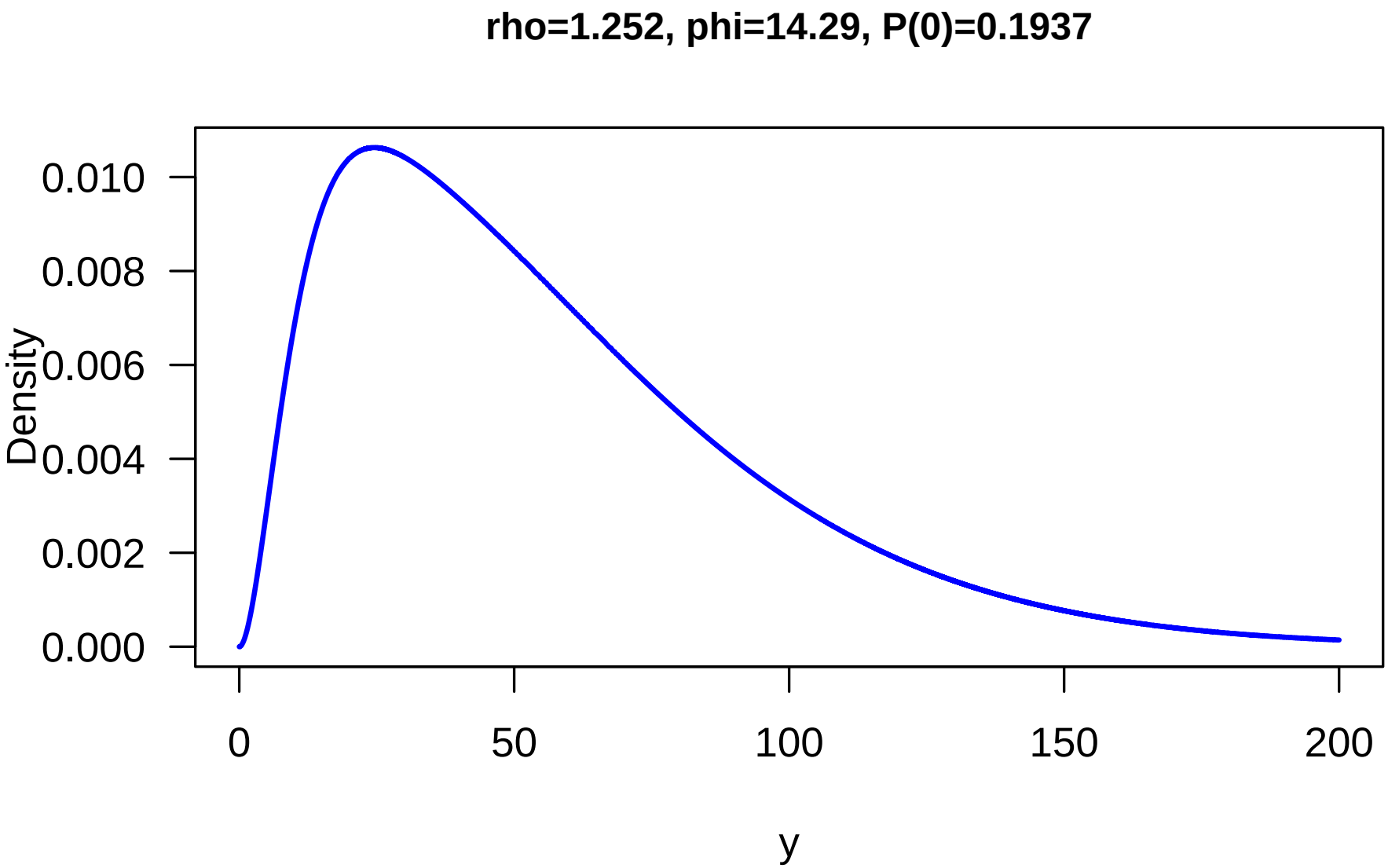




Tweedie examples: $\mu = 10, Var(Y) = 40, \quad 1 < \rho < 2$



Tweedie examples: Catch rate of rays



Comparison of models

We wish to compare two models, lognormal vs Tweedie:

$$y_i + \frac{m^+}{2} \sim LN(\eta_i, \sigma^2)$$

versus

$$y_i \sim Tw(\mu_i, \phi, \rho)$$

Objectives:

1. Compare goodness of fit to iid catch rates
2. Compare goodness of fit to 2-sample data (Expt 1 vs Expt 2)
3. Simulate power and coverage for estimating change in μ from 2-sample data.

Objective 1: AIC to compare iid fits

In the model

$$y_i + \frac{m^+}{2} \sim LN(\eta_i, \sigma^2)$$

m^+ will be treated as a constant for sake of simplicity. Then, the AIC is straightforward to obtain.

The AIC is provided directly from the Tweedie models fitted using glmmTMB or cplm in R.

Objective 1: AIC to compare iid fits

In the model

$$y_i + \frac{m^+}{2} \sim LN(\eta_i, \sigma^2)$$

m^+ will be treated as a constant for sake of simplicity. Then, the AIC is straightforward to obtain.

The AIC is provided directly from the Tweedie models fitted using glmmTMB or cplm in R.

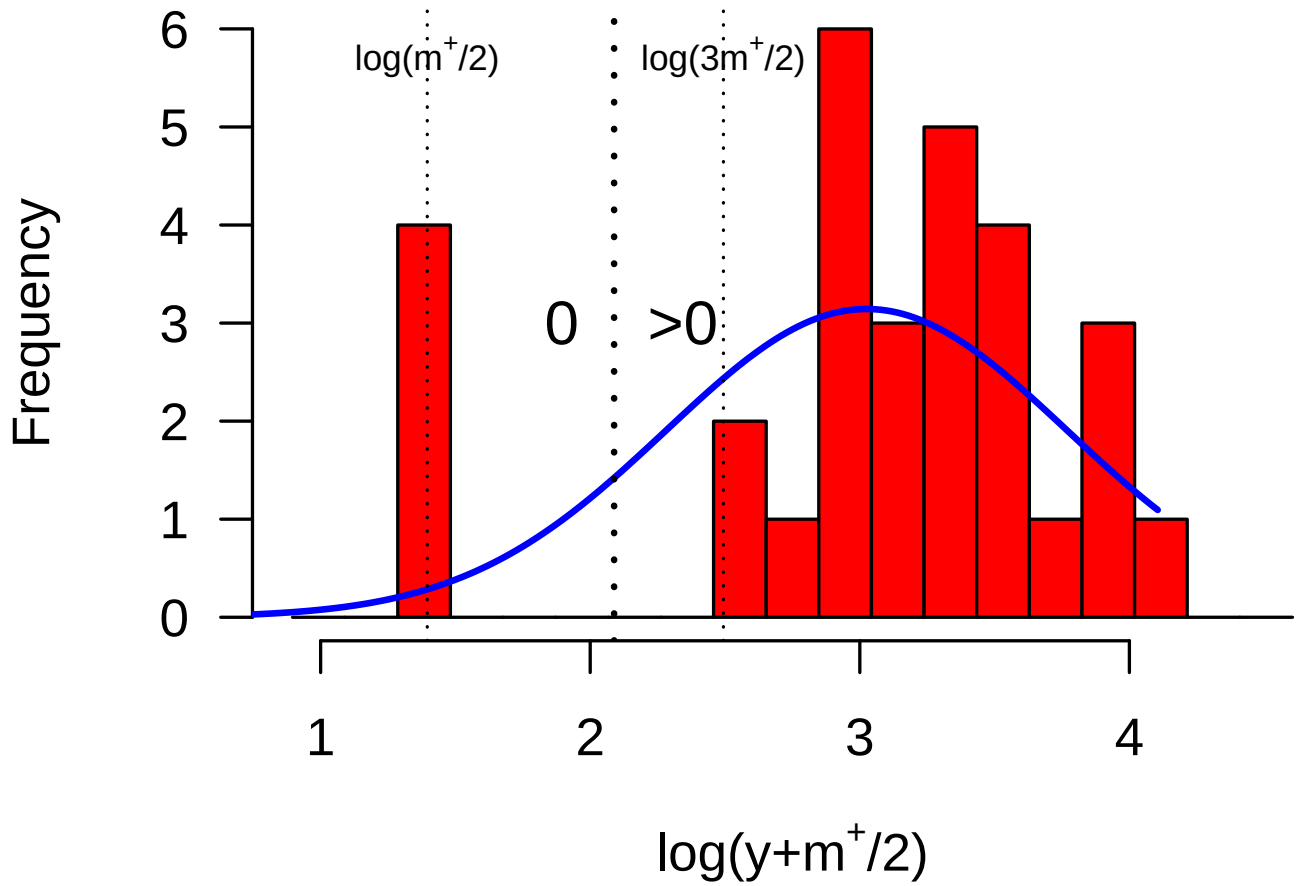
This comparison of AIC **doesn't make sense.** Apples versus oranges since the lognormal is continuous and the Tweedie is a discrete-continous mixture.

Continuous data scale with change of measurement (kg vs tonnes), but zeros don't.

Objective 1: Hybrid AIC to compare iid fits

Although the lognormal is a model for continuous data, for comparison with the Tweedie we need to regard it as a discrete-continuous mixture.

What threshold for zeros and positives to use?



Motivating Example: Catch rate in trawl

Experiment 1 (with horizontal separator panel) and lower codend, AICs:

	Tw	midLN	qtlLN	hiLN	ziLN	ziG	eps	ObsP0	MidLNP0	HiLNP0
Total.com.wt	169.9	169.1	169.1	169.1	169.1	169.9	0.000	0.000	0.000	0.000
Blue.mackere	42.9	45.7	45.8	45.2	44.2	43.2	0.027	0.375	0.231	0.280
Eastern.ange	94.5	103.5	104.2	101.6	93.8	94.5	0.618	0.529	0.338	0.430
Gould.squid.	81.0	82.9	83.1	82.0	77.7	80.4	0.084	0.294	0.197	0.270
John.dory..c	61.2	73.6	74.1	71.7	62.9	61.4	0.050	0.471	0.280	0.358
Morwong..u..	72.2	70.0	70.1	69.8	68.3	71.3	0.027	0.176	0.131	0.181
Redfish..c..	35.9	44.7	46.1	42.8	34.4	35.9	0.087	0.706	0.541	0.691
Red.gurnard.	89.2	99.2	99.4	98.2	93.7	89.9	0.082	0.235	0.133	0.187
Silver.treva	71.0	80.5	81.3	78.2	69.9	71.2	0.112	0.647	0.392	0.467
Tiger.flathe	133.3	138.8	138.2	137.8	137.1	134.2	1.422	0.059	0.018	0.054
Eastern.scho	53.9	68.0	68.7	65.7	55.8	54.1	0.025	0.588	0.336	0.410
Yellowtail.s	43.0	46.5	46.8	45.8	42.9	43.1	0.125	0.400	0.275	0.371
Longfin.bige	20.1	33.3	37.4	37.7	20.0	20.0	0.371	0.882	0.885	0.983
Australian.b	82.7	82.6	82.5	82.4	82.6	82.5	0.476	0.083	0.052	0.106
Blacktip.cuc	129.1	126.3	126.3	126.5	127.0	128.4	0.544	0.083	0.089	0.125
Eastern.fidd	54.9	66.9	69.7	66.7	54.8	54.9	23.753	0.824	0.729	0.889
Gurnard..d..	114.0	115.6	115.6	115.6	115.6	114.0	0.000	0.000	0.000	0.000
Barracouta..	42.7	45.6	45.6	46.0	43.7	43.6	0.279	0.200	0.175	0.298
Velvet.leath	76.6	76.3	76.3	76.2	77.0	76.4	0.250	0.100	0.073	0.116
Smooth.sting	46.0	59.1	61.8	57.4	46.4	45.9	3.807	0.824	0.682	0.843
Misc.rays.sk	159.5	164.5	164.4	164.3	159.3	159.8	4.372	0.176	0.125	0.216
Southern.saw	84.0	86.5	86.7	86.9	79.4	83.9	0.539	0.353	0.304	0.444

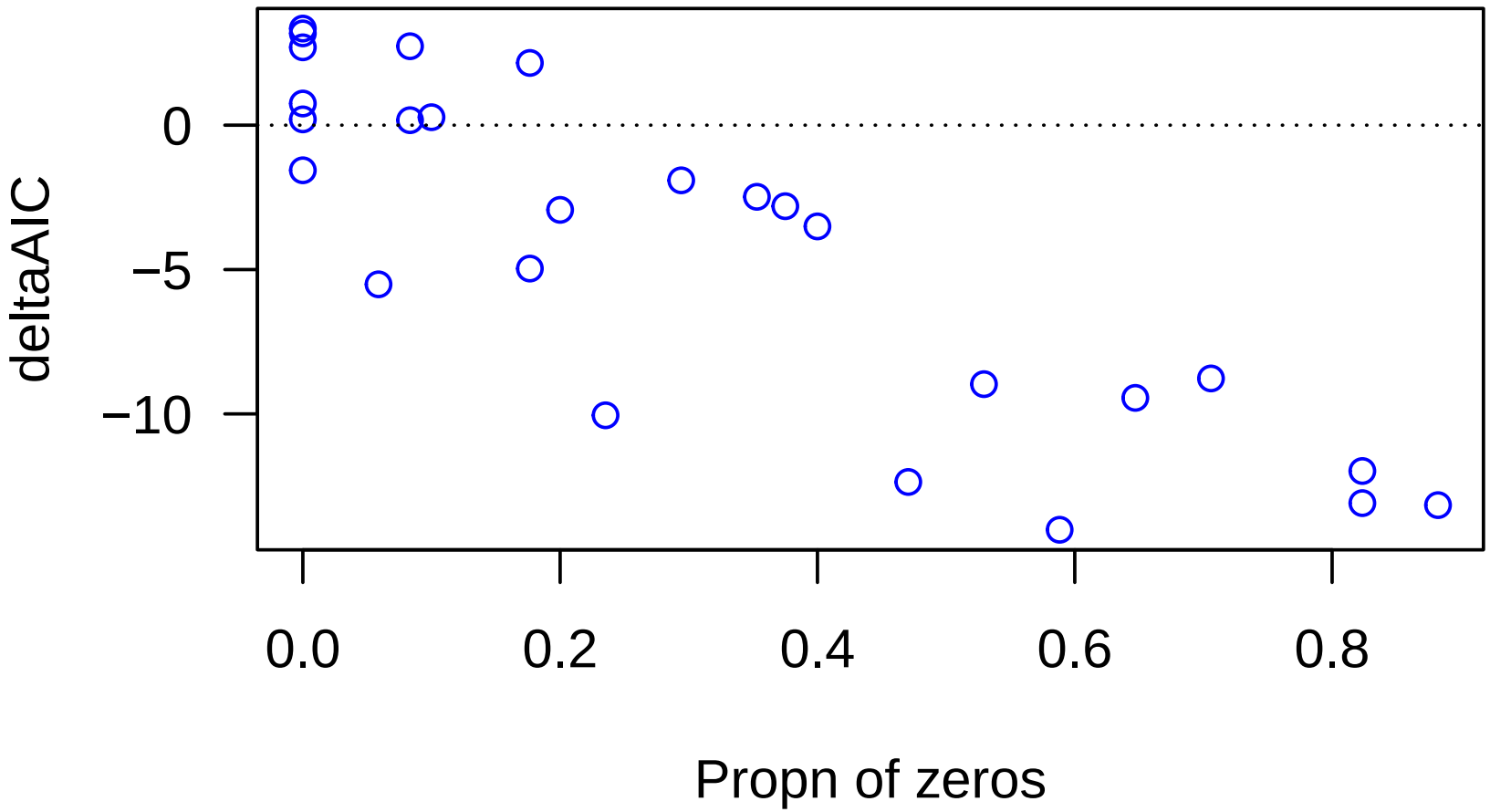
Motivating Example: Catch rate in trawl

Other statistics:

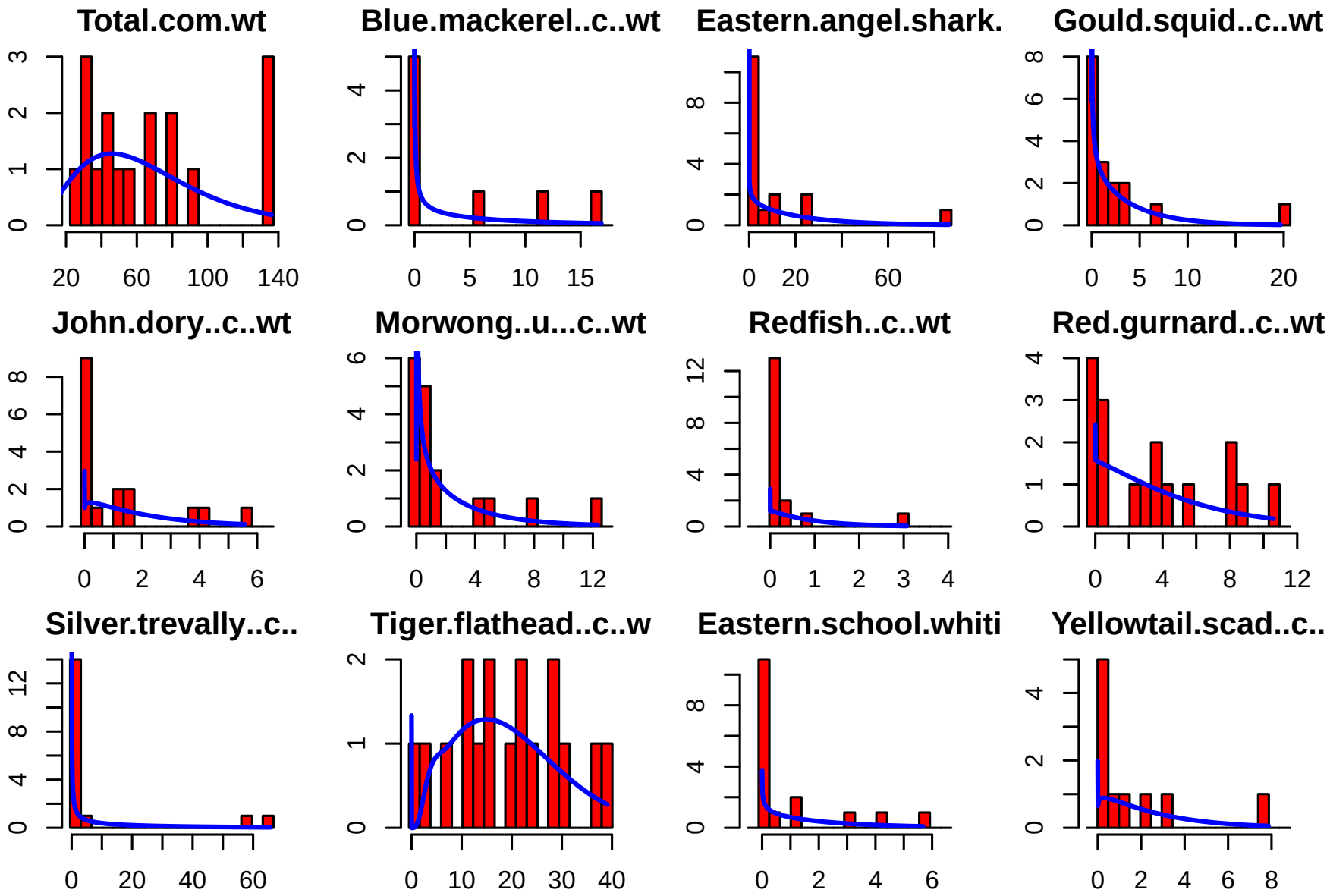
	Tw	midLN	mu	phi	pwr	ObsP0	TwP0	HiLNP0
Total.com.wt	169.9	169.1	66.872	0.314	2.00	0.000	0.000	0.000
Blue.mackere	42.9	45.7	4.235	5.384	1.71	0.375	0.375	0.280
Eastern.ange	94.5	103.5	10.089	10.006	1.55	0.529	0.533	0.430
Gould.squid.	81.0	82.9	2.495	3.180	1.62	0.294	0.311	0.270
John.dory..c	61.2	73.6	1.141	2.701	1.48	0.471	0.469	0.358
Morwong..u..	72.2	70.0	2.040	2.759	1.73	0.176	0.194	0.181
Redfish..c..	35.9	44.7	0.273	2.978	1.50	0.706	0.707	0.691
Red.gurnard.	89.2	99.2	3.462	2.488	1.50	0.235	0.224	0.187
Silver.treva	71.0	80.5	7.625	15.270	1.74	0.647	0.651	0.467
Tiger.flathe	133.3	138.8	19.264	4.262	1.17	0.059	0.037	0.054
Eastern.scho	53.9	68.0	0.918	4.175	1.57	0.588	0.587	0.410
Yellowtail.s	43.0	46.5	1.574	2.651	1.47	0.400	0.403	0.371
Longfin.bige	20.1	33.3	0.131	1.519	1.11	0.882	0.887	0.983
Australian.b	82.7	82.6	8.287	2.520	1.43	0.083	0.098	0.106
Blacktip.cuc	129.1	126.3	85.245	5.168	1.82	0.083	0.092	0.125
Eastern.fidd	54.9	66.9	20.407	66.529	1.24	0.824	0.823	0.889
Gurnard..d..	114.0	115.6	20.353	1.740	2.00	0.000	0.000	0.000
Barracouta..	42.7	45.6	1.860	1.159	1.14	0.200	0.181	0.298
Velvet.leath	76.6	76.3	11.596	3.060	1.64	0.100	0.111	0.116
Smooth.sting	46.0	59.1	4.424	21.203	1.28	0.824	0.826	0.843
Misc.rays.sk	159.5	164.5	46.048	14.289	1.25	0.176	0.193	0.216
Southern.saw	84.0	86.5	3.058	3.287	1.37	0.353	0.377	0.444

Motivating Example: Catch rate in trawl

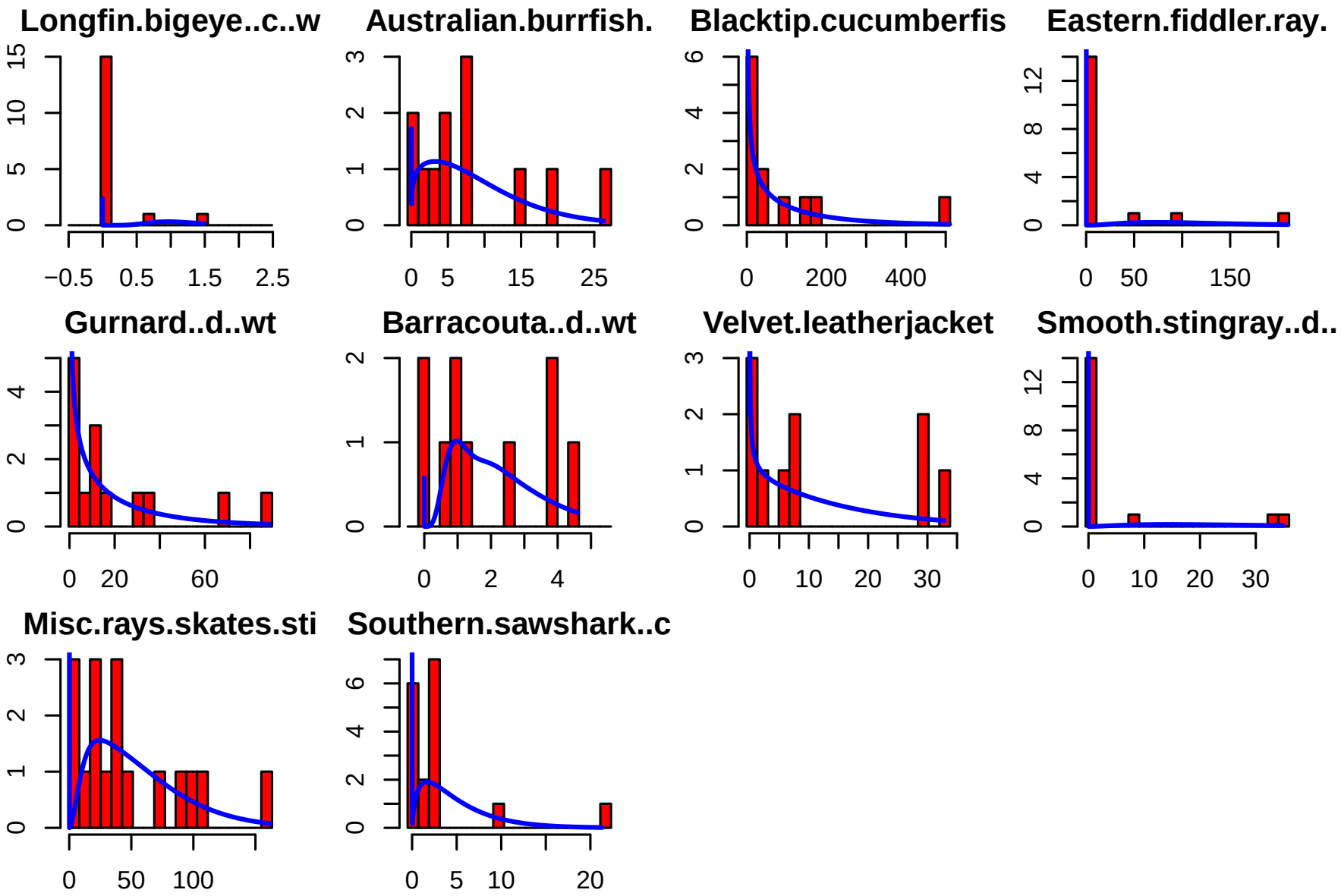
ΔAIC plot: (Tweedie AIC minus midLN AIC) versus proportion of zeros.



Motivating Example: Tweedie fits



Motivating Example: Tweedie fits



Objectives 2 and 3

1. Compare goodness of fit to iid catch rates ✓
2. Compare goodness of fit to 2-sample data (Expt 1 vs Expt 2)
3. Simulate power and coverage for estimating change in μ from 2-sample data.

Objective 3 challenges

- Care is required since the Tweedie model estimates the change in μ, but lognormal approach estimates the change in median.
- `cplm` and `glmmTMB` can give somewhat different fits and quite different estimates of standard errors...**Aghhhh**. Tends to happen when ρ is close to 1.
- Using LRT tests probably safer than Wald tests.

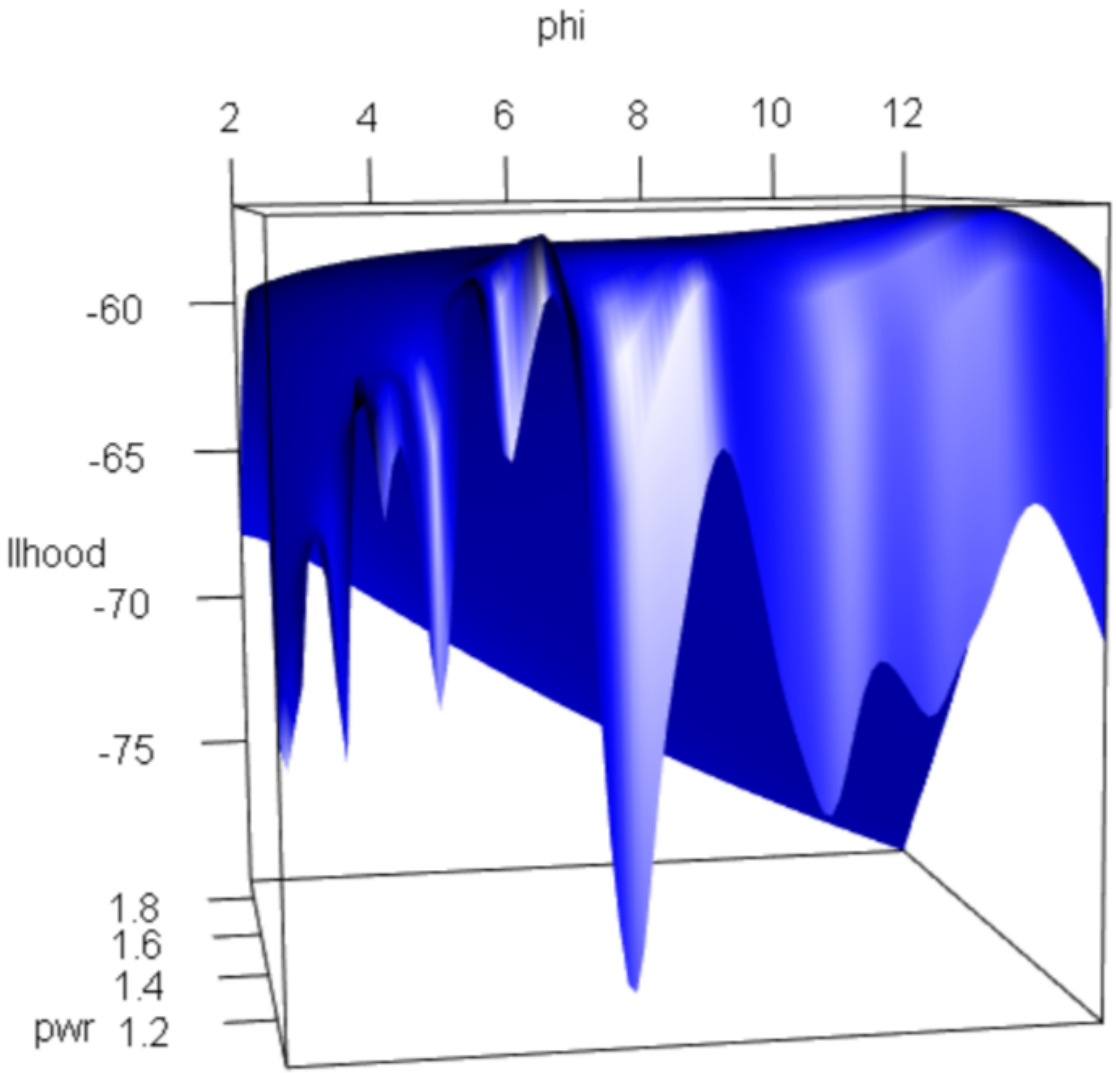
Fit model using optim

```
> #IID Tweedie nllhood
> nllhood1=function(theta,y) {
+   -sum(log( dtweedie(y,mu=theta[1],phi=theta[2],power=theta[3]) )) }
> #With safer parameterization
> nllhood2=function(theta,y,penalty=function(x) 0) {
+   Mu=exp(theta[1])
+   Phi=exp(theta[2])
+   Power=1+plogis(theta[3])
+   -sum(log( dtweedie(y,mu=Mu,phi=Phi,power=Power) ) + penalty(theta)) }
```

```
> optim.fit1=optim(c(mean(y),var(y)/mean(y),1.5),nllhood1,method="L-BFGS-B",
+                  lower=c(mean(y)/1.1,1.01,1.01),upper=c(1.1*mean(y),20,1.99),
+                  hessian=T,y=y,control=list(maxit=1000,trace=0))
>
> optim.fit2=optim(c(log(mean(y)),log(var(y)/mean(y)),0),nllhood2,hessian=T,y=y,
+                   control=list(maxit=1000,trace=0))
```

Convergence issues

Profile log-likelihood of simulated data:



Convergence issues

Profile log-likelihood of simulated data: