How to model just about anything

(Another Boreal Caribou Demography & Ecology Forum)

Elie Gurarie

June 7, 2023

All models have these pieces:

$$Y = f(\mathbf{X} \mid \mathbf{\Theta})$$

- **Y** - response | dependent variable. The thing we want to model / predict / understand. The **effect** (maybe).
- **X** - predictor(s) | independent variable(s) | covariate(s). The thing(s) that "explain(s)" **Y**. The **cause** (maybe).
- **f** - the model structure. This includes: some **deterministic functional form** form (*linear? periodic? polynomial? exponential?*) AND some **probabilistic assumptions**, i.e. a way to characterize the variability / randomness / unpredictability of the process.
- **Θ** - the parameters of the model. There are usually some parameters associated with the **predictors**, and some associated with the **random bit**.

Goals (Art / Science) of Modeling

$$Y = f(\mathbf{X} \mid \mathbf{\Theta})$$

1. Model fitting

What are the **best Θ** values given f, $\mathbf{X}$, Y?

Fitting the model = **estimating the parameters**.

Usually according to some criterion (almost always *Maximum Likelihood*.

2. Model selection

What are the **best** of a set of models f_1, f_2, f_3 given $\mathbf{X}$ and Y?

Different models *usually* vary by what particular variables go into **X**, but can also vary by **functional form** and **distribution assumptions**

Use some **Criterion** (e.g. AIC) to "select" the best model, which balances **how many parameters you estimated** verses **how good the fit is**.

What are "likelihoods"?

Oakie

Orange

Q: What is the "best model" for squirrel morph distribution?

Data and Models

Data / observations: X_{ij}

island	a. Orange	b. Oakie
squirrel 1:	$X_{a,1} = 1$	$X_{b,1} = 1$
squirrel 2:	$X_{a,2} = 1$	$X_{b,2} = 0$

- 1 = light morph
- 0 = dark morph

Models

	model	k
M1:	$P(X_{ij} = 1) = p = 0.5$	1
M2:	$p = 0.75$	1
M3:	$p_a = 1; p_b = .5$	2
M4:	$p_{a,1} = 1;\ \ p_{b,1} = 1 p_{a,2} = 1;\ \ p_{b,2} = 0$ $p_{a,1} = 1;\ \ p_{b,1} = 1$ $p_{a,2} = 1;\ \ p_{b,2} = 0$	4

very important to keep track of the number of parameters!

Likelihoood (of a model)

Product of **probabilities** of **data** given **model**.

$$\mathrm{L}(model) = \prod_{i=1}^{n} \Pr(X \mid model)$$

- We **never** care about the **absolute** value of the likelihood!
- Only the *relative* value of the likelihood.

Four different Squirrel Models:

Data:

$$X_{a1} = 1;\ X_{a2} = 1; X_{b1} = 1;\ X_{b2} = 0$$

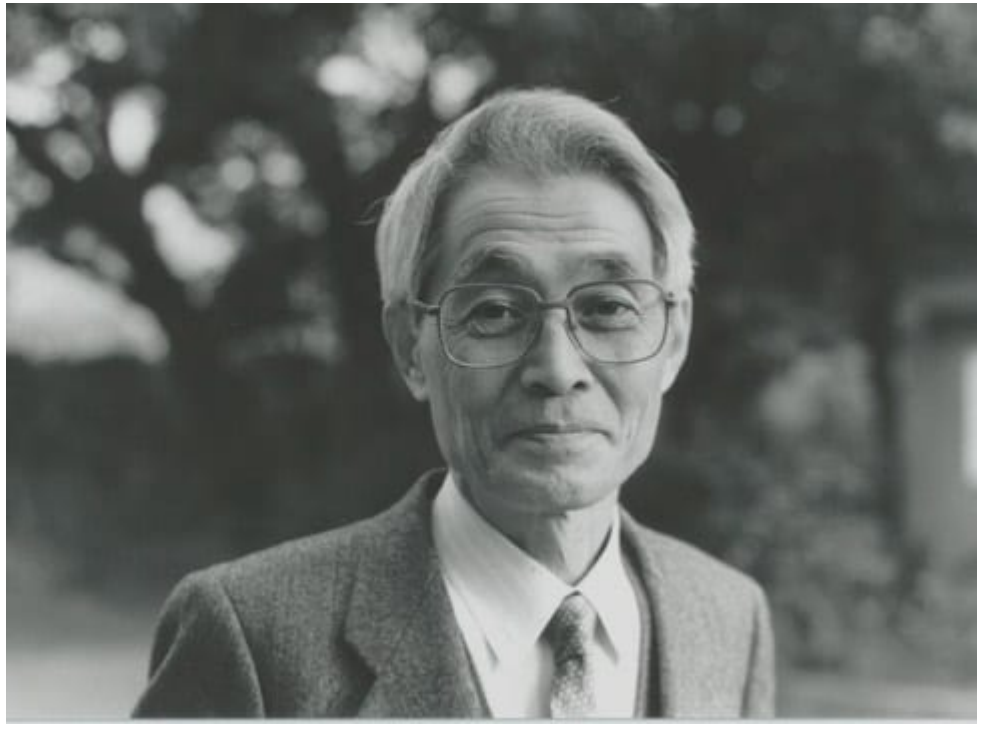
model		likelihood	
M1	$p = 0.5$	$\frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2}$	0.0625
M2	$p = 0.75$	$\frac{3}{4} \times \frac{3}{4} \times \frac{3}{4} \times \frac{1}{4}$	0.1055
M3	$p_a = .5; p_b = 1$	$1 \times 1 \times \frac{1}{2} \times \frac{1}{2}$	0.25
M4	$p_{a,1} = 1;\ p_{b,1} = 1;\ p_{a,2} = 1;\ p_{b,2} = 0$	$1 \times 1 \times 1 \times 1$	1

$$L(M4) > L(M3) > L(M2) > L(M1)$$

*M4** has the higest likelihood! But is this a useful model?

A(kaike) Information Criterion

A good fit is great! But it is useless if it uses too much information (too many parameters). This is *overfitting*.
One parameter per data point is TOO MANY parameters!



Hirotugo Akaike 赤池 弘次 (1927-2006)

Simple formula:

$$AIC = -2\log(\mathrm{L}) + 2k$$

(where k is the number of parameters)

- Better fit = higher L = lower AIC.
- Too complicated = more k = higher AIC.

Lowest AIC is "best" model

Compute AIC

model	likelihood	log-likelihood	k	AIC
M1: coin flip	0.0625	-2.77	1	7.55
M2: proportional odds	0.1055	-2.25	1	**6.50**
M3: island specific	0.25	-1.39	2	6.77
M4: individual specific	1	0	4	8

AIC2 < AIC3 < AIC1 < AIC4

Most *parsimonious* model is M2!

Conclusion: not enough evidence to identify a difference between islands.

Let's add one more observation ...

Oakie Island

Orange Island

$$X_{b,3} = dark$$

Updated squirrel models:

model	probs	Likelihood		k	AIC
M1	$p = \frac{1}{2}$	$\frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2}$	= 0.03125	1	8.93
M2	$p = \frac{3}{4}$	$\frac{3}{4} \times \frac{3}{4} \times \frac{3}{4} \times \frac{1}{4} \times \frac{1}{4}$	= 0.02637	1	9.27
M2b	$p = \frac{3}{5}$	$\frac{3}{5} \times \frac{3}{5} \times \frac{3}{5} \times \frac{2}{5} \times \frac{2}{5}$	= 0.0346	1	8.73
M3	$p_a = .5; p_b = 1$	$1 \times 1 \times \frac{1}{2} \times \frac{1}{2} \times 0$	= 0 (!!)	2	∞
M3b	$p_a = \frac{1}{2}; p_b = \frac{2}{3}$	$\frac{1}{2} \times \frac{1}{2} \times \frac{2}{3} \times \frac{2}{3} \times \frac{1}{3}$	= 0.037	2	10.6

Updated (1 parameter) squirrel models:

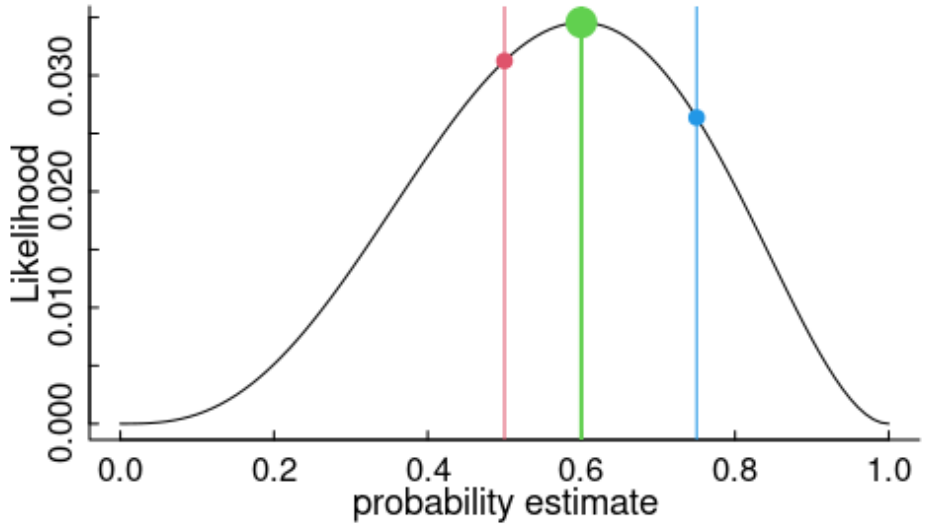
model	probs		L
M1	$p = \frac{1}{2}$	$\frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2}$	0.03125
M2	$p = \frac{3}{4}$	$\frac{3}{4} \times \frac{3}{4} \times \frac{3}{4} \times \frac{1}{4} \times \frac{1}{4}$	0.02637
M2b	$p = \frac{3}{5}$	$\frac{3}{5} \times \frac{3}{5} \times \frac{3}{5} \times \frac{2}{5} \times \frac{2}{5}$	0.0346

If you sweep through all possible values of p, you find that $\hat{p} = 3/5$ leads to the highest likelihood.

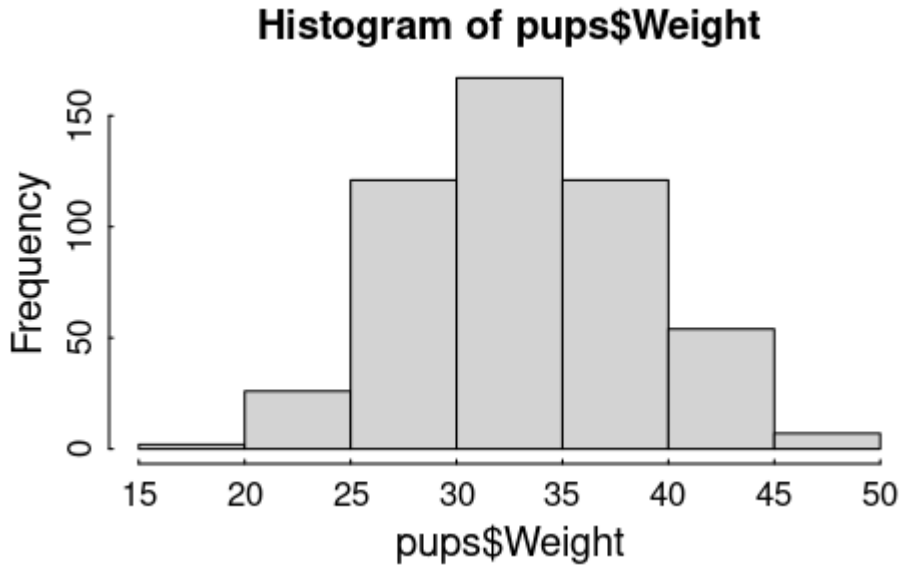
This is the **maximum likelihood estimate** (MLE) of the probability that a squirrel is light morph.

But you can also get (good) Confidence Intervals from looking at the curve of the profile.

Likelihood profile



Null (linear) model




```
mean(pups$Weight)
```

```
## [1] 33.51004
```

```
sd(pups$Weight)
```

```
## [1] 5.661695
```

This suggests a model!

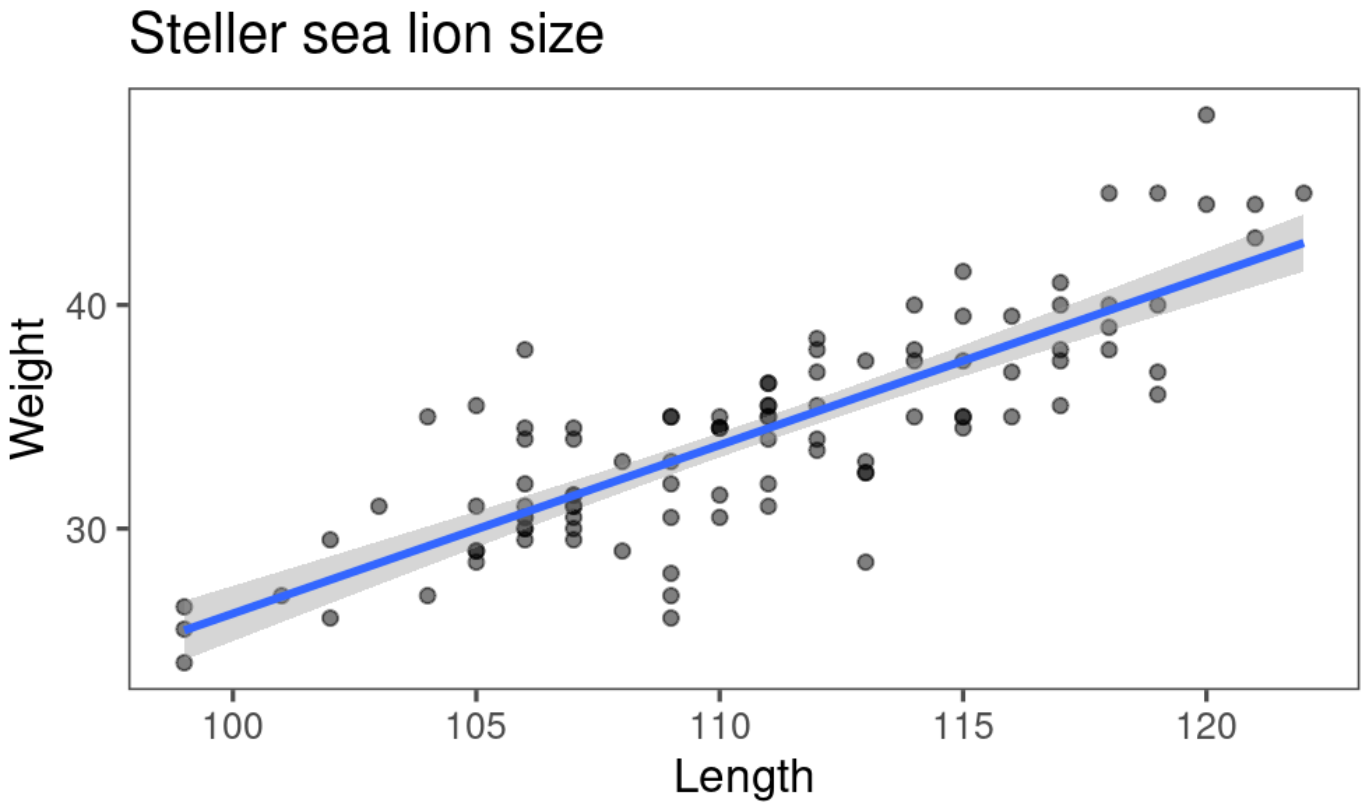
$$W \sim \mathrm{N}(\mu = 33\,kg, \sigma = 5.7)$$

With no covariates.

Simple linear model

Probably there is a relationship between length and weight. The simplest relationship is linear.

$$Y \sim \text{N}(\text{mean} = \beta_0 + \beta_1 X,\ \text{sd} = \sigma)$$



Steller sea lion (*Eumatopias jubatus*) pups.

Deterministic model:

$$Y_i = \beta_0 + \beta_1 X_i$$

- β_0 - intercept
- β_1 - slope

This is the **functional form of the predictor**

Statistical model:

Version 1:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

where $\epsilon_i \sim \mathrm{N}(0, \sigma)$

or

Version 2:

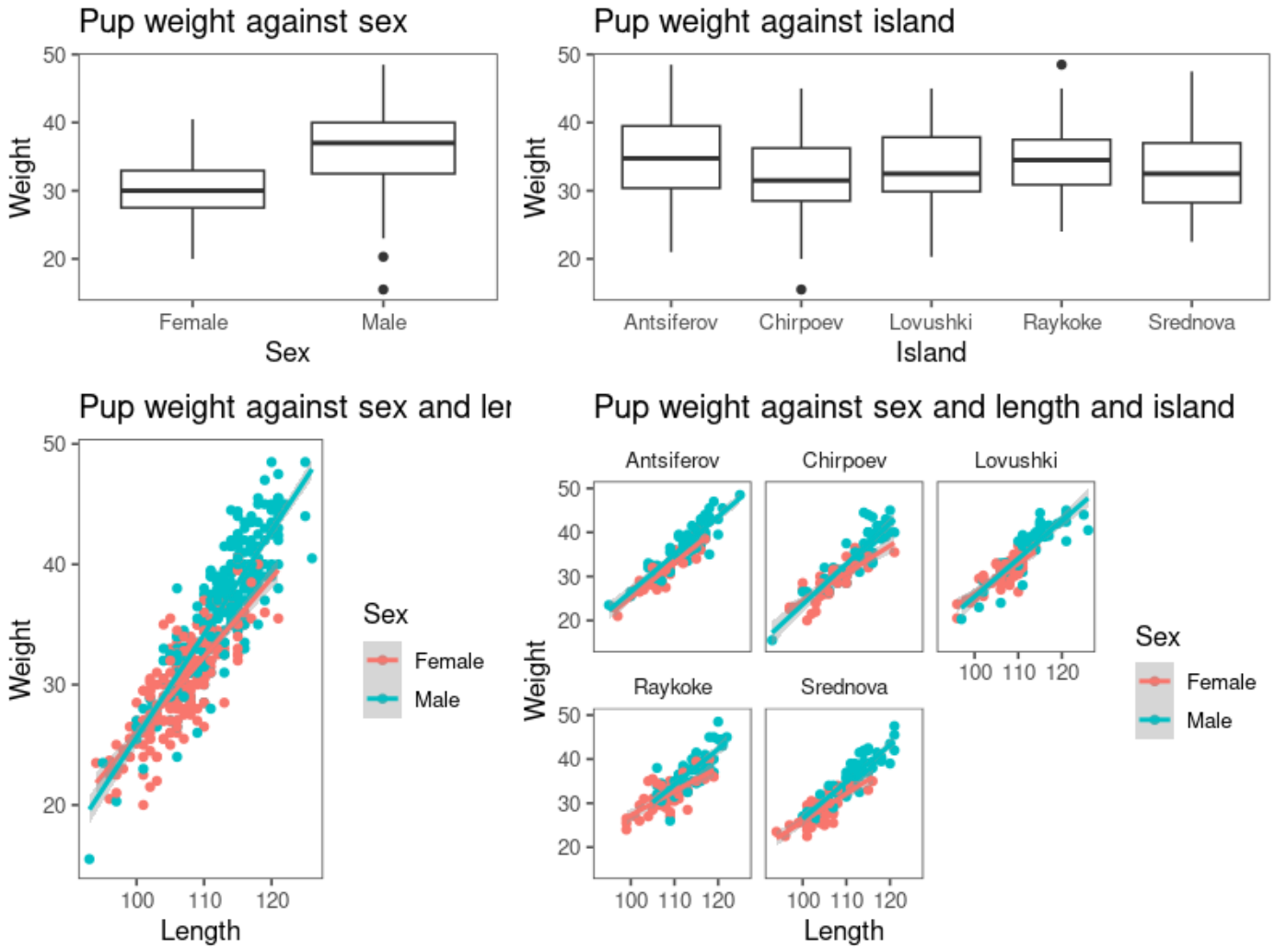
$$Y_i \sim \mathrm{N}(\beta_0 + \beta_1 X_i, \sigma)$$

V2 is better because it is more transparent about the number of parameters!

- Two (intercept | slope) are part of the **functional form**
- One (residual standard deviation) is part of the **random component**.

But other variables might influence pup size

Lots of competing models with different **main** and **interaction** effects.



Fitting and model selection

Model	k	R2	logLik	AIC	dAIC
Weight ~ Length * Sex + Island	**8**	**0.818**	**-1144.6**	**2307.3**	**0.0**
Weight ~ Length * Sex * Island	20	0.824	-1137.1	2316.1	8.8
Weight ~ Length + Sex + Island	7	0.811	-1155.0	2325.9	18.6
Weight ~ Length * Sex	4	0.803	-1164.5	2339.0	31.7
Weight ~ Length + Sex	3	0.795	-1174.5	2357.0	49.7
Weight ~ Length	2	0.779	-1193.4	2392.8	85.5
Weight ~ Sex	2	0.293	-1483.2	2972.4	665.1
Weight ~ Island	5	0.028	-1562.5	3137.0	829.7
Weight ~ 1	1	0.000	-1569.5	3143.1	835.8

This is what we expect ... the interaction between **sex** and **length** is consistent across islands, but there are some main effect differences across islands (mainly because of the time we sampled).

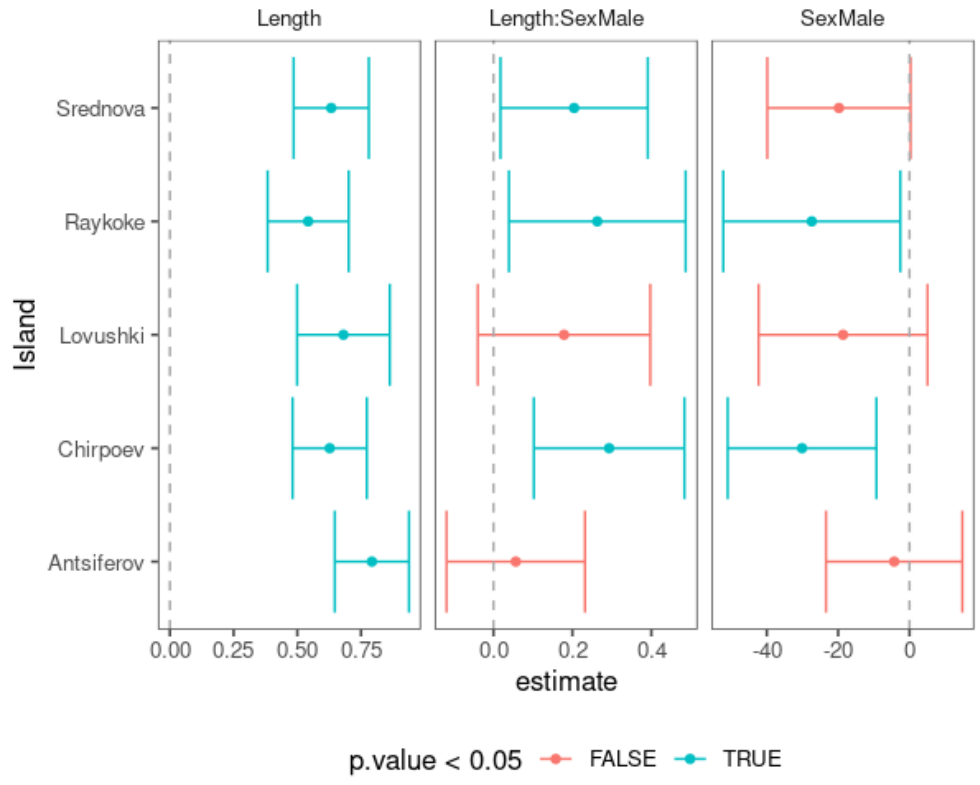
Model selection vs. parameter estimates

The best model:

$$Y_{ijk} = \beta_0 + \beta_{1i}\text{Island}_{ijk} + (\beta_2 + \beta_{3j}\text{Sex}_{ijk}) \times (\text{Length}_{ijk}) + \epsilon_{ijk}$$

What are the **parameter estimates** (effect sizes) of the selected model?

term	estimate	std.error	statistic	p.value
SexFemale	-39.34	3.69	-10.67	0.00
SexMale	-59.29	3.06	-19.37	0.00
Length	0.66	0.03	19.11	0.00
IslandChirpoev	-2.00	0.34	-5.81	0.00
IslandLovushki	-0.47	0.34	-1.35	0.18
IslandRaykoke	-0.45	0.35	-1.31	0.19
IslandSrednova	-0.35	0.35	-1.00	0.32
SexMale:Length	0.20	0.04	4.55	0.00

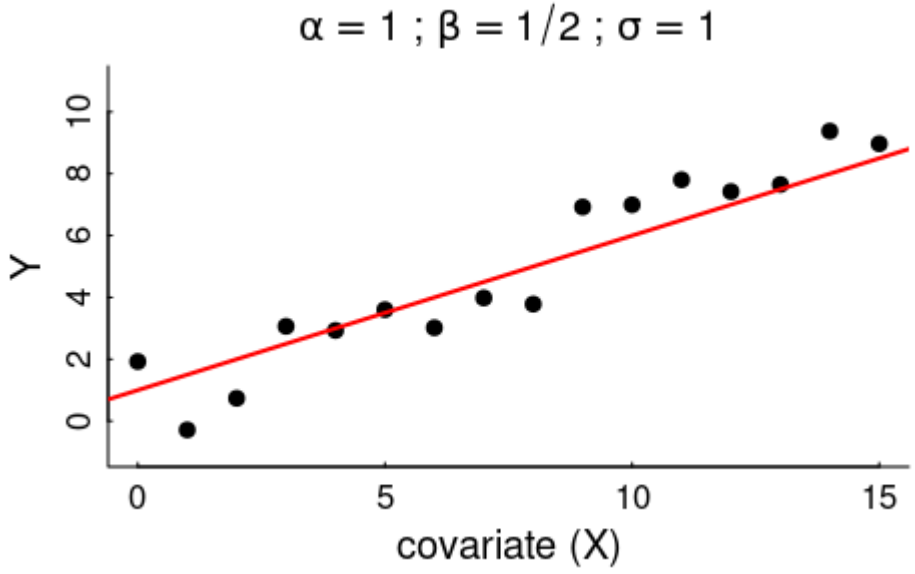


Generalized linear model

Normal Model

$$Y_i \sim \text{Normal}(\alpha_0 + \beta_1 X_i, \sigma)$$

Models continuous data with a "normal-like" distribution.

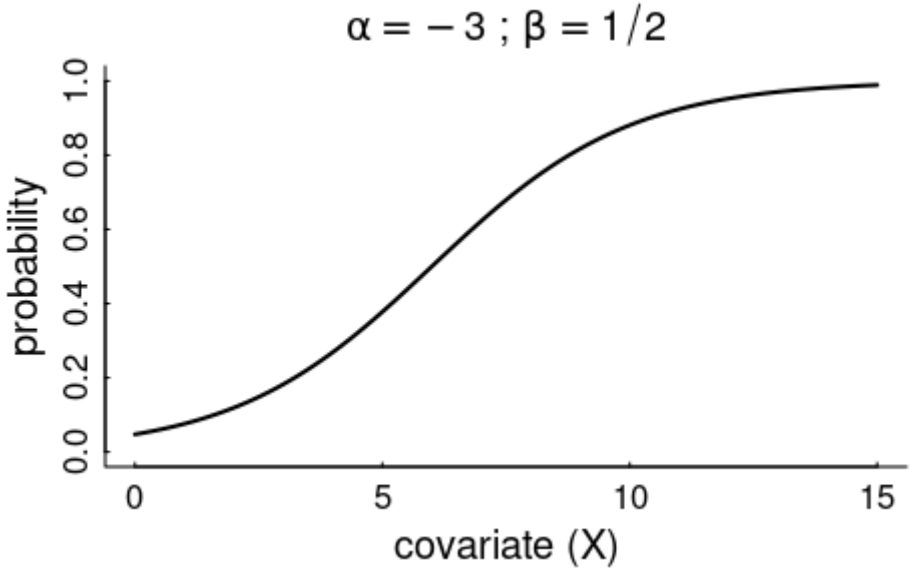


Binomial model

$$Y_i \sim \text{Bernoulli}\left(\frac{\exp(\alpha + \beta X_i)}{1 + \exp(\alpha + \beta X_i)}\right)$$

There's some *probability* of something happening that depends on the predictor X.

Bernoulli just means the data are all 0 or 1.

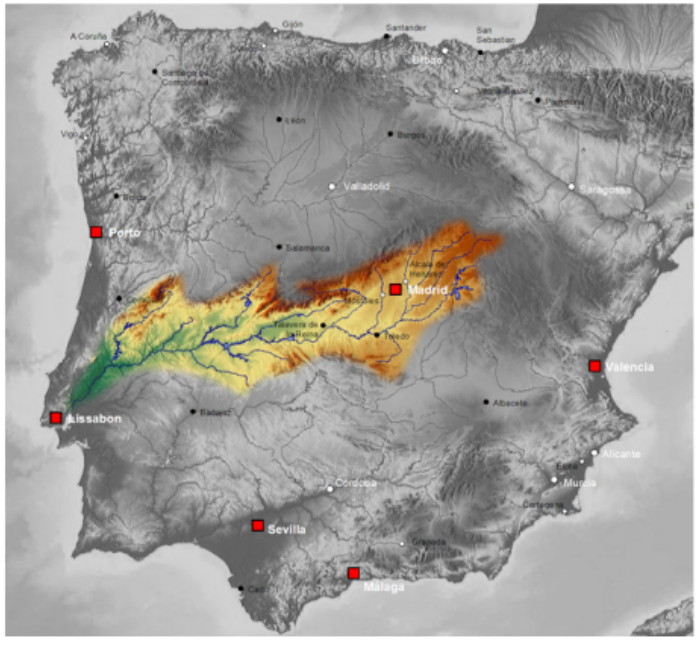


This models **presence/absence**, **dead/alive**, **male/female** other response variables with **2** possible outcomes.

What factors predict occurence of *Solea solea* larvae?

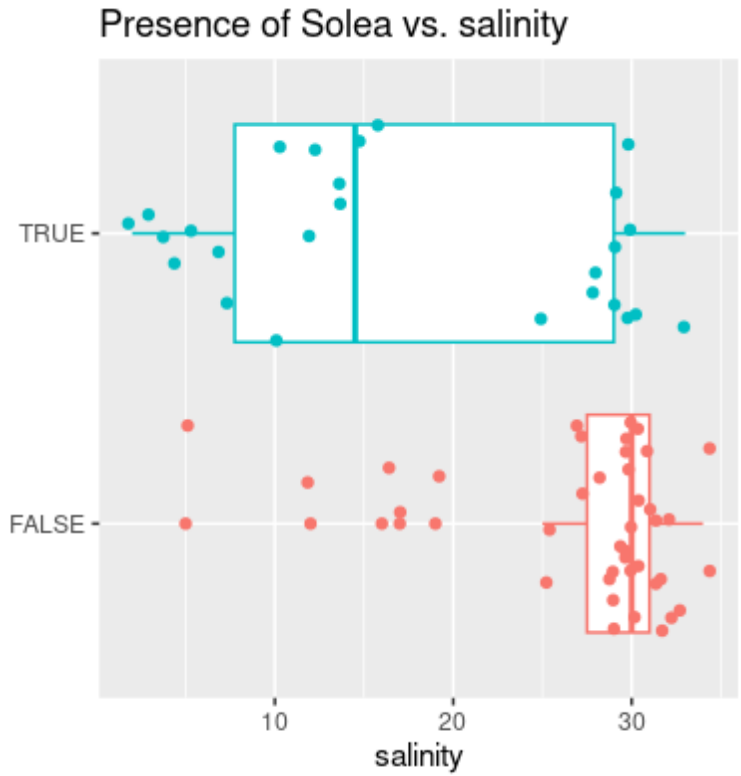
Sampled in the estuary of the Tejo river in Portugal

- Lots of environmental factors in data



depth	temp	salinity	transp	gravel	large_sand	fine_sand	mud	presence
3.0	20	30	15	3.74	13.15	11.93	71.18	**0**
2.6	18	29	15	1.94	4.99	5.43	87.63	**0**
2.6	19	30	15	2.88	8.98	16.85	71.29	**1**
2.1	20	29	15	11.06	11.96	21.95	55.03	**0**
3.2	20	30	15	9.87	28.60	19.49	42.04	**0**
3.5	20	32	7	32.45	7.39	9.43	50.72	**0**

Presence of *Solea solea* against **salinity**



Modeling is EXACTLY the same as **linear regression** except:

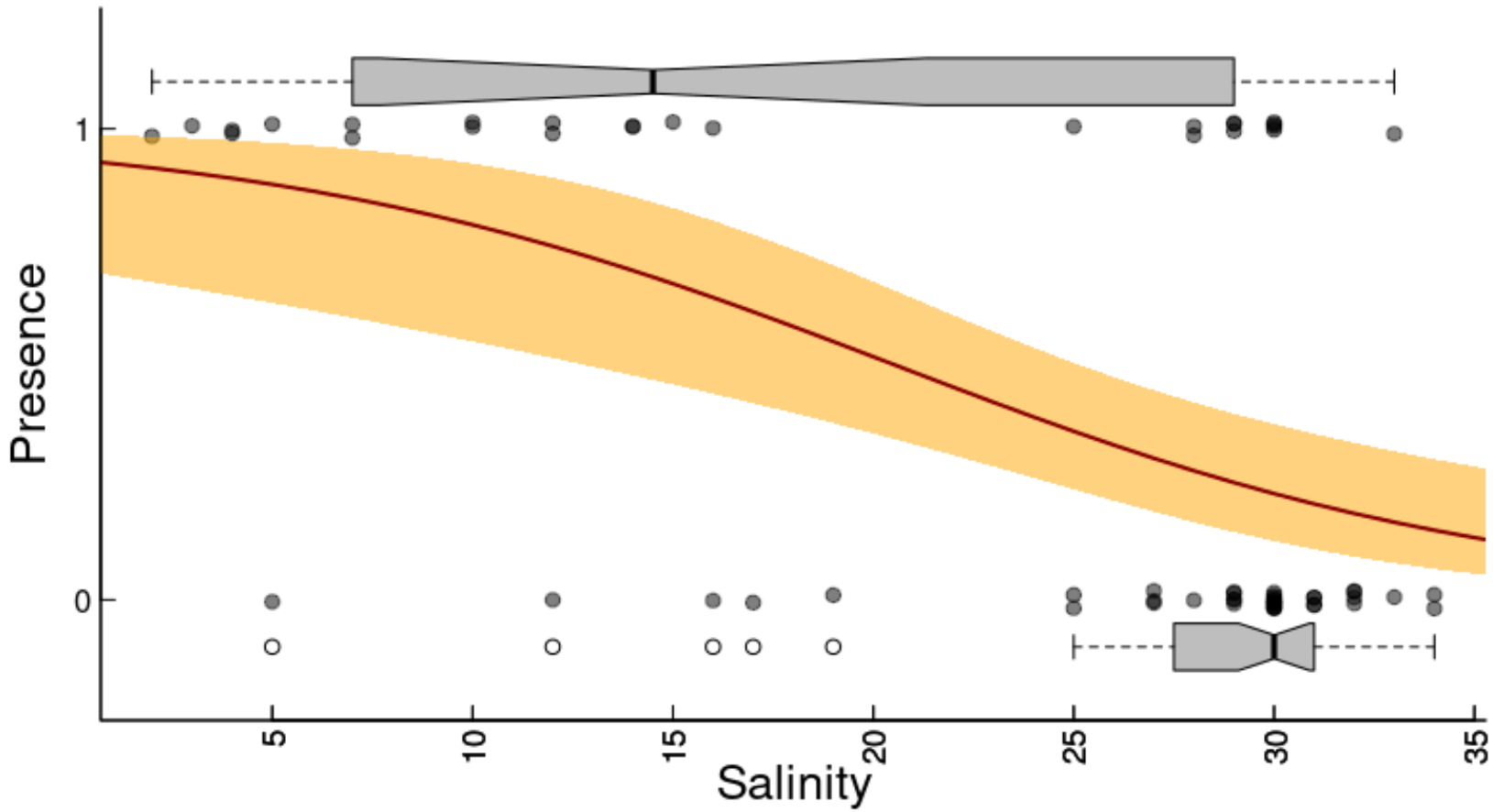
- `glm` - for **generalized** linear model (instead of `lm`)
- `family = 'binomial'` is the instruction to fit the logistic regression

```
glm(presence ~ salinity, family ='binomial')
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	2.661	0.902	2.951	0.003
salinity	-0.130	0.035	-3.716	0.000

Clearly - *Solea solea* presence is very significantly *negatively* related to salinity.

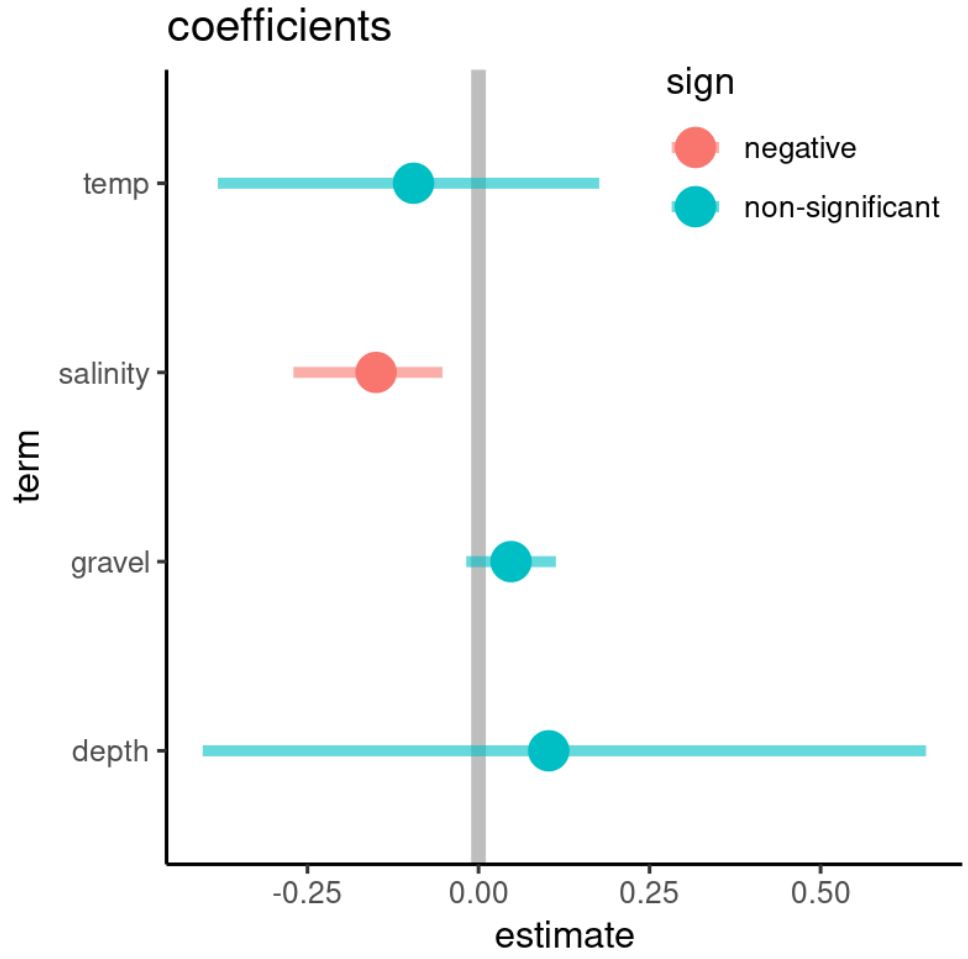
Out of this model we can make predictions



ΔAIC analysis - and coefficients

	Model	k	logLik	AIC	dAIC
M9	**salinity + gravel**	**3**	**-33.2**	**72.5**	**0.0**
M2	**salinity**	**2**	**-34.3**	**72.6**	**0.1**
M7	**temp + salinity**	**3**	**-34.0**	**74.0**	**1.5**
M5	**depth + salinity**	**3**	**-34.1**	**74.3**	**1.8**
M11	depth + temp + salinity	4	-33.9	75.8	3.3
M0	depth	2	-38.1	80.1	7.6
M4	depth + temp	3	-38.0	81.9	9.4
M6	depth + gravel	3	-38.0	82.0	9.5
M10	depth + temp + gravel	4	-37.8	83.7	11.2
M1	temp	2	-43.3	90.6	18.1
M3	gravel	2	-43.7	91.3	18.8
M8	temp + gravel	3	-43.3	92.6	20.1

Salinity clearly among the more important covariates (in the top 4 models).



Take-aways on (linear) modeling

1. **Statistical modeling** separates **patterns** (the model) from "**randomness**" (unexplained variation).

2. We structure our models to have a **response variable** and one or more **predictors** or **covariates**.

3. Depending on the response variable, a different **family** is chosen:

   - if **continuous** and symmatric: **Normal**
   - if 1/0 (presence/absence, dead/alive): **Binomial**
   - if counts: **Poisson**
   - if lots of zeros: **zero-inflated** or **hurdle**

4. An important task is **Model selection**, identifying which model is "best"

   - Best means *"explains the most variation without overfitting"*
   - Very common criterion is **AIC.**
   - lots of **back and forth** between **fitting** and **selection**

5. Once a model is "selected", we can:

   - analyze the results by seeing the **effect sizes** (magnitude of coefficients, aka *slopes*) and **directions** (signs of coefficients)
   - make **inferential predictions** by "spreading" our model over a larger landscape.

6. **Well over 90% of wildlife modeling (esp. habitat) is done along these lines!**

Funkier (non-linear) - but relevant model

Received: 4 July 2019 | Accepted: 12 September 2019

DOI: 10.1111/2041-210X.13305

APPLICATION

Methods in Ecology and Evolution | BRITISH ECOLOGICAL SOCIETY

For everything there is a season: Analysing periodic mortality patterns with the `cyclomort` R package

Eliezer Gurarie[1] | Peter R. Thompson[1,2] | Allicia P. Kelly[3] | Nicholas C. Larter[4] | William F. Fagan[1] | Kyle Joly[5]

Under the hood:
Maximum Likelihood Estimation

mortalities probability distribution:

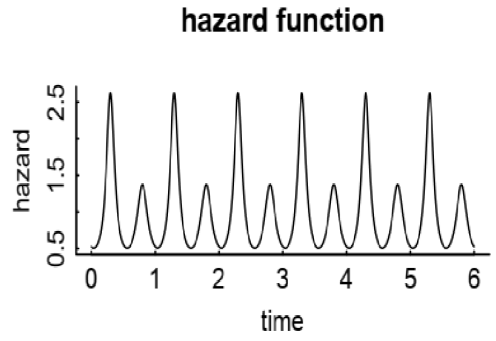
$$f(t|\theta) = h(t|\theta)\exp(-\int_{t=0}^{t} h(t'|\theta)dt')$$

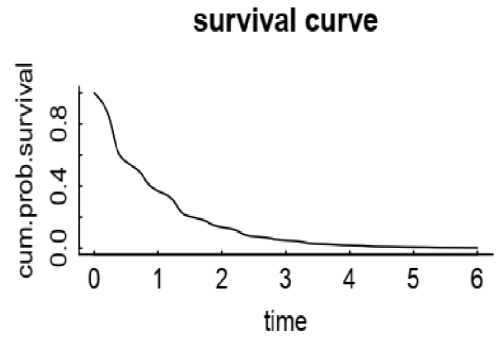
```
T.morts.sim <- simulate_cycloSurv(300, period = 1, peaks = c(0.3, 0.8),
                                  durations = c(0.15, 0.20), weights = c(0.6,0.4),
                                  censoring = "none", meanhazard = 1,
                                  plotme = TRUE, max.periods = 6)
```

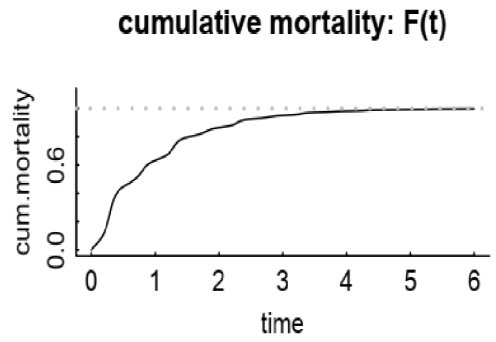
mortalities likelihood:

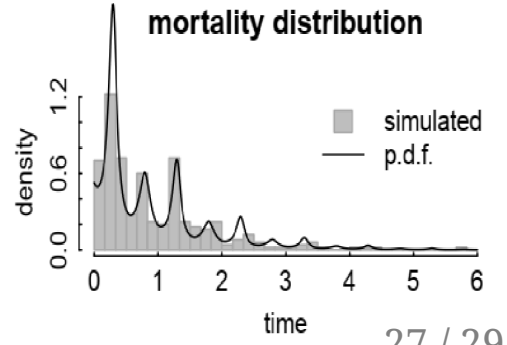
$$L(\theta|T_i, T_{0,i}) = \prod_{i=1}^{n} h(T_i|\theta)\exp\left(-\int_{T_{0,i}}^{T_i} h(t'|\theta)dt'\right)$$

Note: *this is the Likelihood for* ***any*** *time-to-event process [only slightly more complicated if censored]*

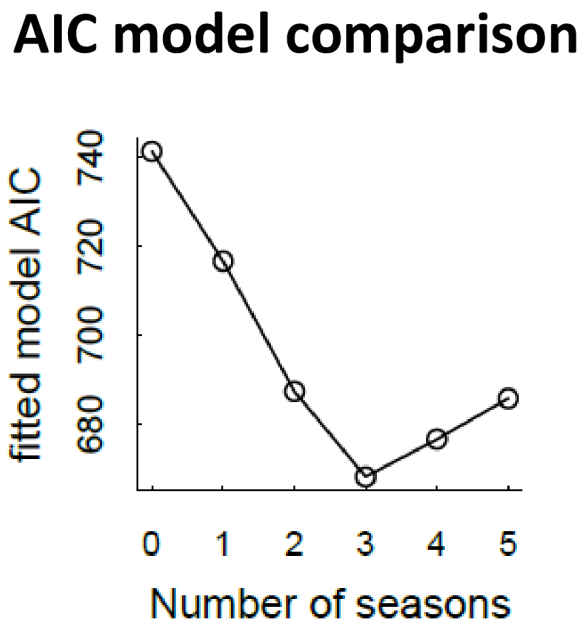


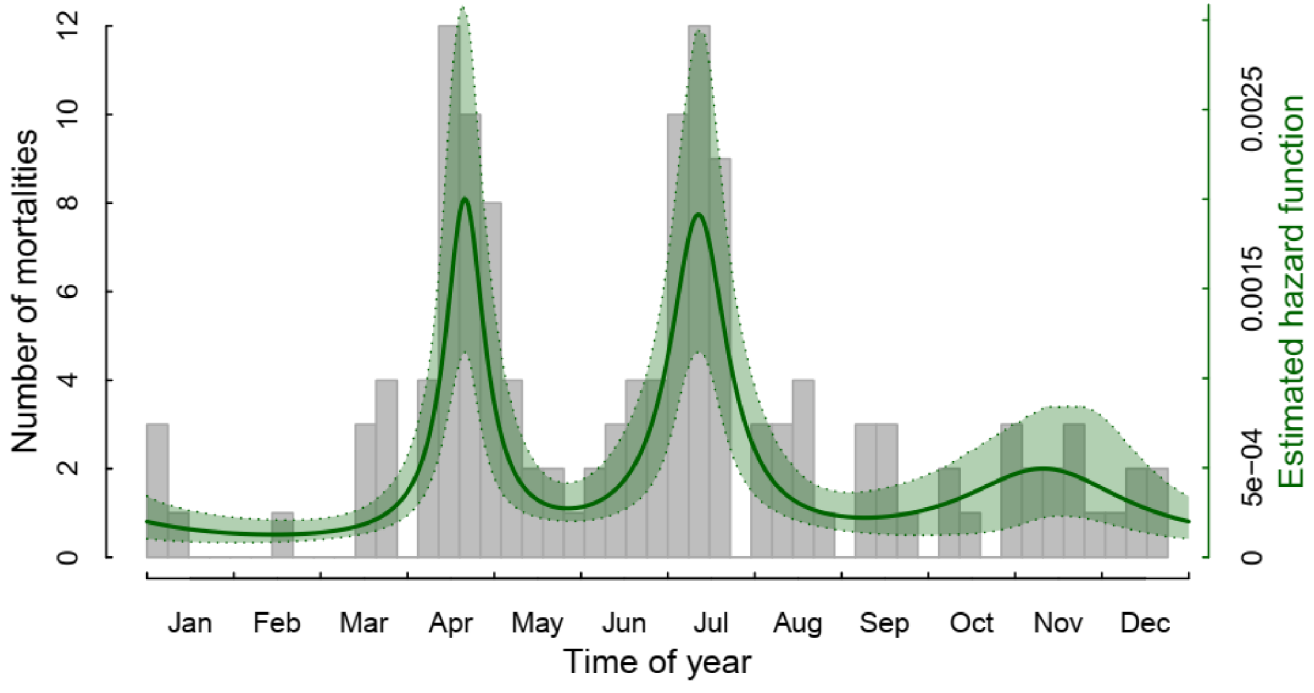






NWT: estimating the 3-humped camel





	Season 1		Season 2		Season 3	
Parameter	**Estimate**	**(95% CI)**	**Estimate**	**(95% CI)**	**Estimate**	**(95% CI)**
Peak	Apr 21	(Apr 17–Apr 26)	Jul 13	(Jul 8–Jul 18)	Nov 8	(Oct 11–Dec 6)
Duration (days)	17.74	(10.3–29.7)	23.58	(12.5–42.2)	71.07	(26.7–129)
Weight	0.30	(0.2–0.4)	0.40	(0.2–0.6)	0.30	(0.1–0.5)
Mean hazard rate ($year^{-1}$)	0.17	(0.16 – 0.19)	–	–	–	–

Take home message

- Not all models have to be linear! You can get creative and try to capture a process of actual biological interest.
- As long as you can **write the model** (i.e. the likelihood function), you can *probably* **fit the model**, and just remember to **keep track of the number of parameters**!