BIOL 501: Model Selection

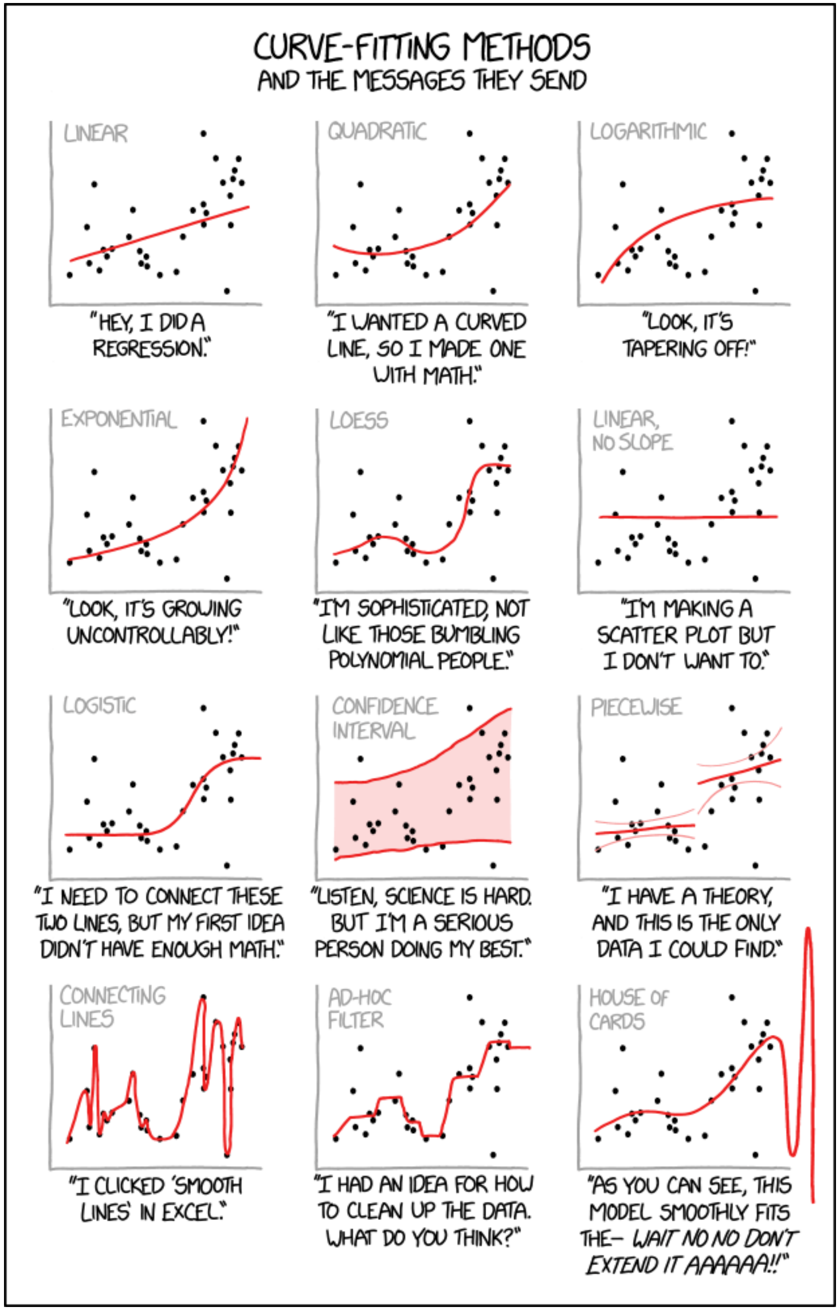
1\.

Peer-Feedback Survey

2\.

Canvas announcement jamboard link next slide

3\.


CURVE-FITTING METHODS
AND THE MESSAGES THEY SEND
LINEAR
"HEY, I DID A REGRESSION."
QUADRATIC
"I WANTED A CURVED LINE, SO I MADE ONE WITH MATH."
LOGARITHMIC
"LOOK, IT'S TAPERING OFF!"
EXPONENTIAL
"LOOK, IT'S GROWING UNCONTROLLABLY!"
LOESS
"I'M SOPHISTICATED, NOT LIKE THOSE BUMBLING POLYNOMIAL PEOPLE."
LINEAR, NO SLOPE
"I'M MAKING A SCATTER PLOT BUT I DON'T WANT TO."
LOGISTIC
"I NEED TO CONNECT THESE TWO LINES, BUT MY FIRST IDEA DIDN'T HAVE ENOUGH MATH."
CONFIDENCE INTERVAL
"LISTEN, SCIENCE IS HARD. BUT I'M A SERIOUS PERSON DOING MY BEST."
PIECEWISE
"I HAVE A THEORY, AND THIS IS THE ONLY DATA I COULD FIND."
CONNECTING LINES
"I CLICKED 'SMOOTH LINES' IN EXCEL."
AD-HOC FILTER
"I HAD AN IDEA FOR HOW TO CLEAN UP THE DATA. WHAT DO YOU THINK?"
HOUSE OF CARDS
"AS YOU CAN SEE, THIS MODEL SMOOTHLY FITS THE- WAIT NO NO DON'T EXTEND IT AAAAAA!!"

7 March 2023 Dr. Beth Volpov, b.volpov@oceans.ubc.ca; UBC 2022 Term 2

xkcd.com/2048

Office Hours Announcements

- **Beth this week**
  - **Wednesday March 8th: 3:30-4:30pm** (Zoom link on left tab of Canvas home page)
  - **Thursday March 9th: 12-1pm** (in person) AERL #245

- **Avery this week**
  - **Tuesday (today)** 3-4pm Tuesdays (in person, DMP #101)
- **Avery *next* week**
  - Tuesday 3-4pm (DMP #101)
  - Wednesday 1-2pm (BRC #336)
  - Friday 1-2pm (BRC #336)

Can also email **either of us** for an appointment. Suggest 3 specific dates/times and we will pick one

Assignment #2

Jamboard Anonymous Poll

What is the status of your Assignment #2?

- **Red/pink sticky**= I haven’t found a dataset yet
- **Yellow sticky**=I have a dataset and know if I’m using glm, lm, or lme
- **Green sticky=**I have a dataset, know model type I am doing, and have started model analysis in R

https://jamboard.google.com/d/1sCCkvR55TNRlOSleHw_MGmKcEFG-fxiKxgE-0pEQtdc/edit?usp=sharing

Assignment #2: Due 18 March @9pm

- **Linear, mixed, or generalized linear model in R**. You can choose which type of **linear** model to use, as long as it is appropriate for your dataset.
  - Fixed→Lm, glm, gam
  - Mixed→LME, GLMM, GAMM
- Only 1 response variable
- At least 1 categorical factor
- Include at least 1, and no more than 2, additional explanatory variables

Self-assessment with rubric *prior* to turning it in or do a peer-review

How to find data for Assignment #2

- Ask labmates, supervisor, other grad students for data
- Extract data that was used for something else other than a linear model
  - "How to find data for practicing R.docx"/ Assignment #2 on Canvas
- **Not** ok to redo a figure that's already published and analyzed as a linear model, but you can extract data from published paper that was used for something else

- **Don't** use simulated data
- Yes it's ok, if it's your own thesis data, undergrad data

Find your dataset asap. Ask us if you have questions sooner rather than later

Outline

- The problem of model selection
- Goals of model selection
- AIC Criterion
- Forwards vs backwards selection strategies
- Search strategies: dredge() and stepAIC()
- Several models may fit equally well
- The Science Part: formulate a set of candidate models
- Review model selection with linear models (full vs reduced)
- Workshop Prep

“Essentially, all models are wrong, but some models are useful.”

George E.P. Box

All models have error.

Models are simplified representations with assumptions.

You can make a line for any model, but that doesn’t mean that specific line is a good fit (or the only fit).

All models have error.

Take Home Point

It’s a grey area—there is no “right” answer

- **Be transparent** on the model selection process, packages used, and why parameters were included or excluded
- Your model selection and choice of parameters should **make biological sense with thought ahead of time on which parameters are meaningful**
- Think about experimental design ahead of time—can you minimize confounding factors
- Data analysis should follow your question design and experimental design

You have already done model selection with LME models

LME model (X,Y, both numerical)

1. Fit null (reduced model)
   - null.model<-lme(y~x,random=~x|animal, data=mydata)
   - Assess model fit-**"Is it linear"?**
     - anova(null)
2. Fit full (+factor1 model) ***that only varies by adding this 1 fixed factor***
   - model1<-lme(y~x+**fixed factor1**,random=~x|animal, data=mydata)
   - Assess model fit- **"Is it linear"?**
     - anova(model1)
3. Compare hierarchically nested models with LRT test and choose lowest AIC if p is sig
   - Anova(null,model1) – **"Which model is better?" or "Does adding this fixed factor improve the model better than the null model?"**

Reduced vs full models

- **Nested models:***Reduced vs. full models are referred to as “nested models”, because the one contains a subset of the terms occurring in the other.
- **Non-nested models:** Models in which the terms contained in one are **not** a subset of the terms in the other are called “non-nested” models.
- **Don’t confuse this with nested experimental designs or nested sampling designs.

Example of model comparison with lm (fixed effects only)

#1.scatter plot (examine data)

```
plot(y ~ x, data = mydata)
```

#2. Fit linear model

```
model1<- lm(y ~ x, data=mydata)
```

#3. Extract coefficients and information from the model

summary(model1) and model1$coefficients

#4.Add model line to scatter plot above

abline() or lines() or ggplot()

Plot CI with visreg()

predict()

#5. Test model fit with anova

```
anova(model1)
```

#6. Look at model assumptions (diagnostics)

```
plot(model1)
```

#7. Predict() new data from model line (in workshop)

Goals of model selection

- A model that predicts (from new data) well
  - Cross-validation is one option
- A model that approximates the ‘true’ relationship between the variables.

Goal is to balance goodness of fit with simplicity

- If a model includes **too many** predictors→common issue could be overfitting
  - Gives good predictions to training data but poor predictions when applied to new data model not trained on
  - Low bias but high variance
- If a model includes **too few** predictors→common issue could be underfitting
  - Gives poor predictions
  - Low variance but high bias

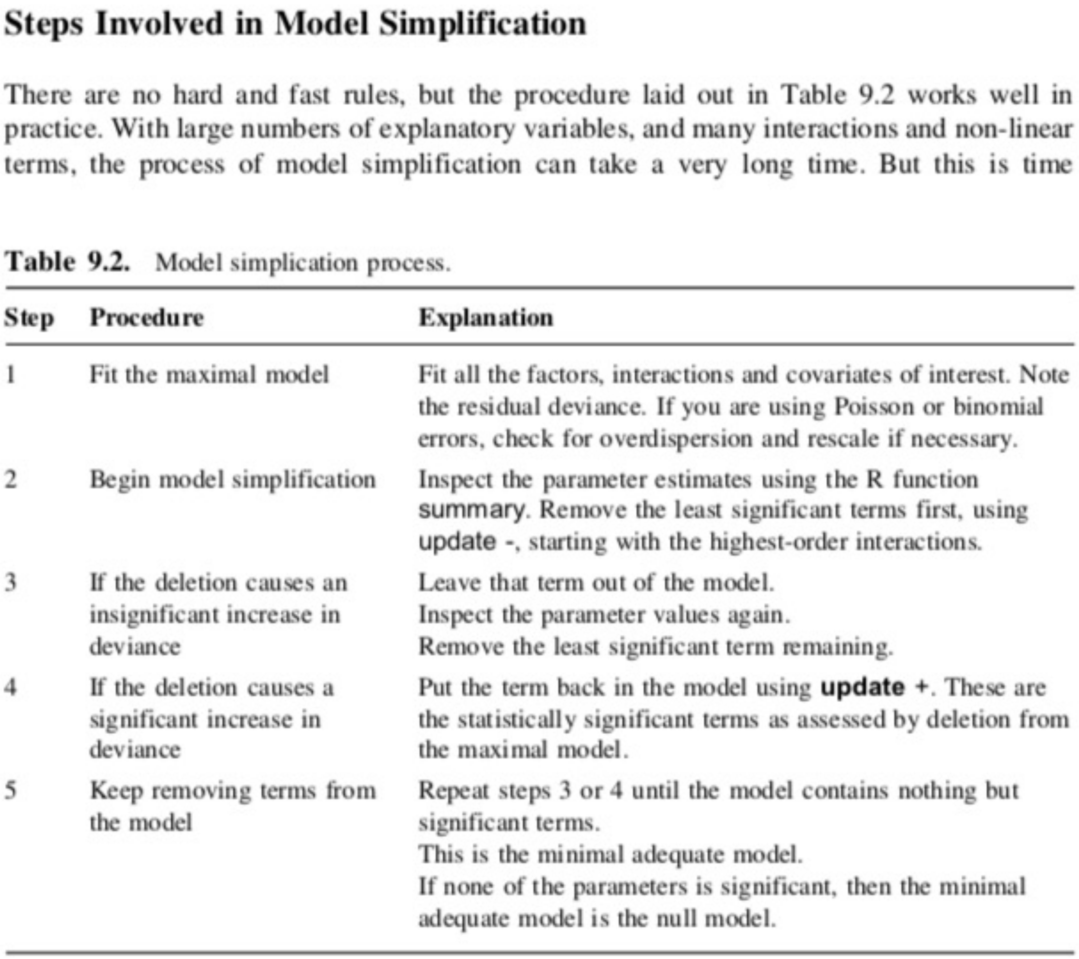
The problem of model selection

- **Parsimony principle:** Fit no more parameters than is necessary. If two or more models fit the data almost equally well, prefer the simpler model.

- *“models should be pared down until they are minimal adequate”*
  -- Crawley 2007, p325

Additional resources

Crawley R Book: Ch 9 Statistical Modelling



Steps Involved in Model Simplification

There are no hard and fast rules, but the procedure laid out in Table 9.2 works well in practice. With large numbers of explanatory variables, and many interactions and non-linear terms, the process of model simplification can take a very long time. But this is time

Table 9.2. Model simplication process.

Step	Procedure	Explanation
1	Fit the maximal model	Fit all the factors, interactions and covariates of interest. Note the residual deviance. If you are using Poisson or binomial errors, check for overdispersion and rescale if necessary.
2	Begin model simplification	Inspect the parameter estimates using the R function summary. Remove the least significant terms first, using update -, starting with the highest-order interactions.
3	If the deletion causes an insignificant increase in deviance	Leave that term out of the model. Inspect the parameter values again. Remove the least significant term remaining.
4	If the deletion causes a significant increase in deviance	Put the term back in the model using **update +**. These are the statistically significant terms as assessed by deletion from the maximal model.
5	Keep removing terms from the model	Repeat steps 3 or 4 until the model contains nothing but significant terms. This is the minimal adequate model. If none of the parameters is significant, then the minimal adequate model is the null model.

The problem of model selection: Stepwise multiple regression

- Using stepwise elimination or addition of terms, is a common practice
- Fitting a multiple regression with many variables, cycle of adding/deleting model terms and then refitting.
- Continue until only statistically significant terms remain

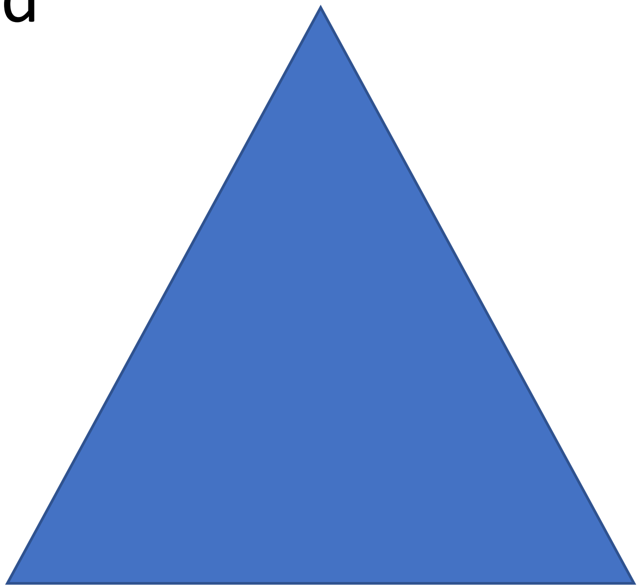
The problem of model selection: Stepwise multiple regression

- Stepwise multiple regression yields a **single, final model**, the “minimum adequate model.”(MAM) or “best fit model”

- But is this a good idea? Does it really yield the best model? (Discussion and paper today)

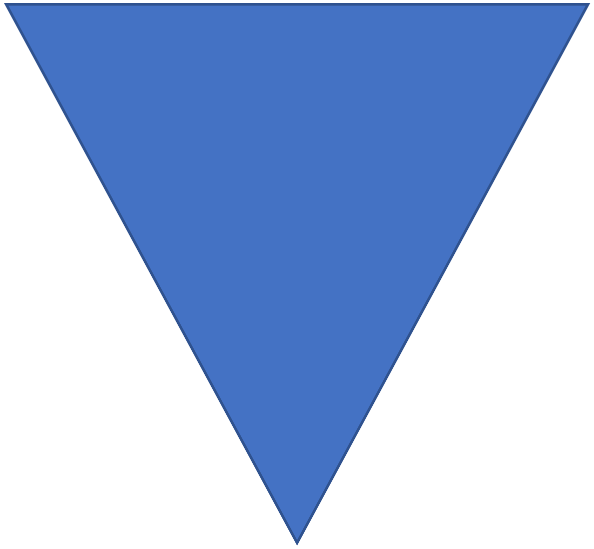
Forwards stepwise selection

- Start with **null** model containing no predictors (reduced or ‘empty model’)
- **Add** significant predictors to the model, one-at-a-time
- Each cycle compare full vs. reduced



Backward stepwise selection

- **Start with full model** containing all predictors (all parameters, variables included)
- **Remove nonsignificant** predictors to the model, one-at-a-time
- Each cycle compare full vs. reduced
- Dredging is an example



Backwards vs. Forwards Model Selection

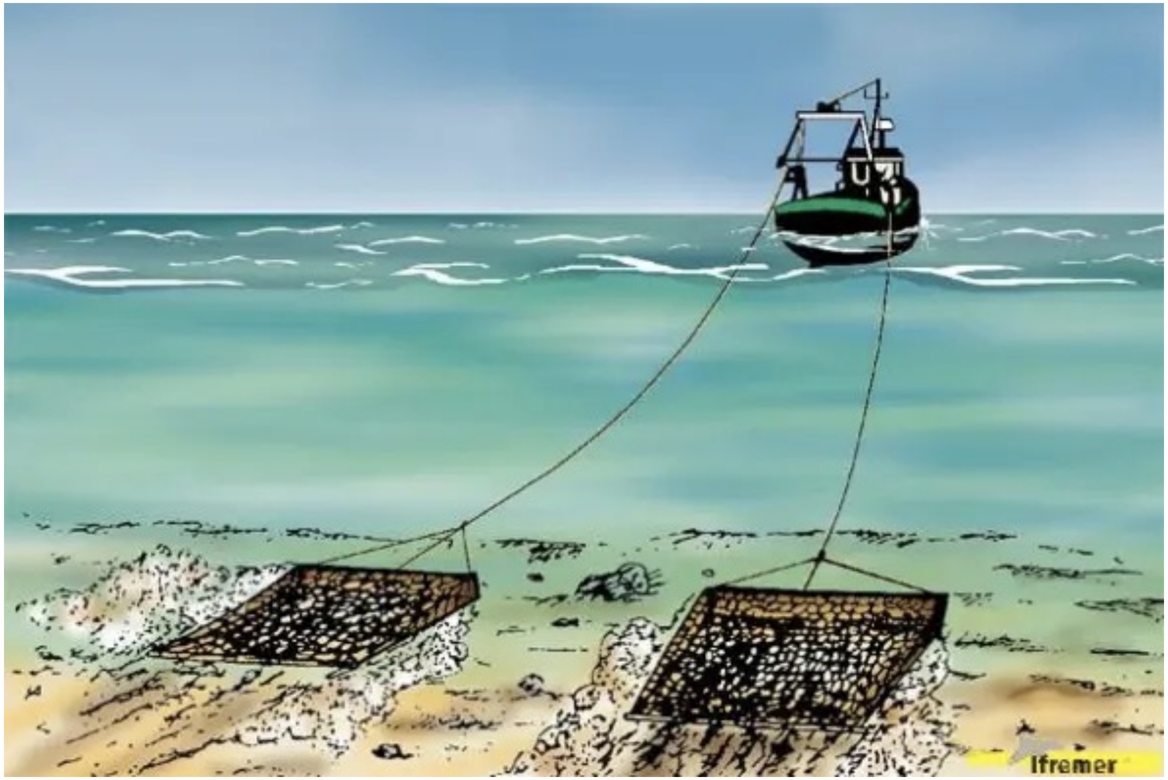
- Some people prefer to use backwards model selection for “exploratory” analysis and forwards model building when you have more prior knowledge on what may or may not be a factor
- Some disciplines have standard ways of model building
- **But it’s a grey area—very few hard and fast rules across multiple disciplines**

Does stepwise elimination of terms actually yield the "best" model?

- What criterion are we actually using to decide which model is "best"?
- Each step in which a variable is dropped from the model involves "accepting" a null hypothesis.
    - What happens if we drop a false null hypothesis?
    - How can a sequence of Type 2 errors lead us to the "best" model?
- How repeatable is the outcome of stepwise regression? With a different sample, would stepwise elimination bring us to the same model again?
- Might models with different subsets of variables fit the data nearly as well?

Data dredging

What is dredging on a boat?



Pull up everything in the net and let's see what we get

What is dredging data? (backwards model selection)

- Extensive automated search
- Not have any hypotheses to help guide search and model building
- Is there any good, a priori reason to a term/model among the set of candidate models to evaluate?

Next: Example of combining data dredging with cross-validation

P-hacking usually refers to the dragging of statistical significance out of data related to one or more hypotheses of interest, data dredging is the extensive search for significant relationships in a dataset without necessarily having specific hypothesis in mind

Example of combining data dredging and backwards model selection with cross-validation

Can we predict foraging success in fur seals from a wide variety of TDR variables?

- GLMM/GAMM analysis with animal random factor
- Utilize historical TDR databases
- Dredged

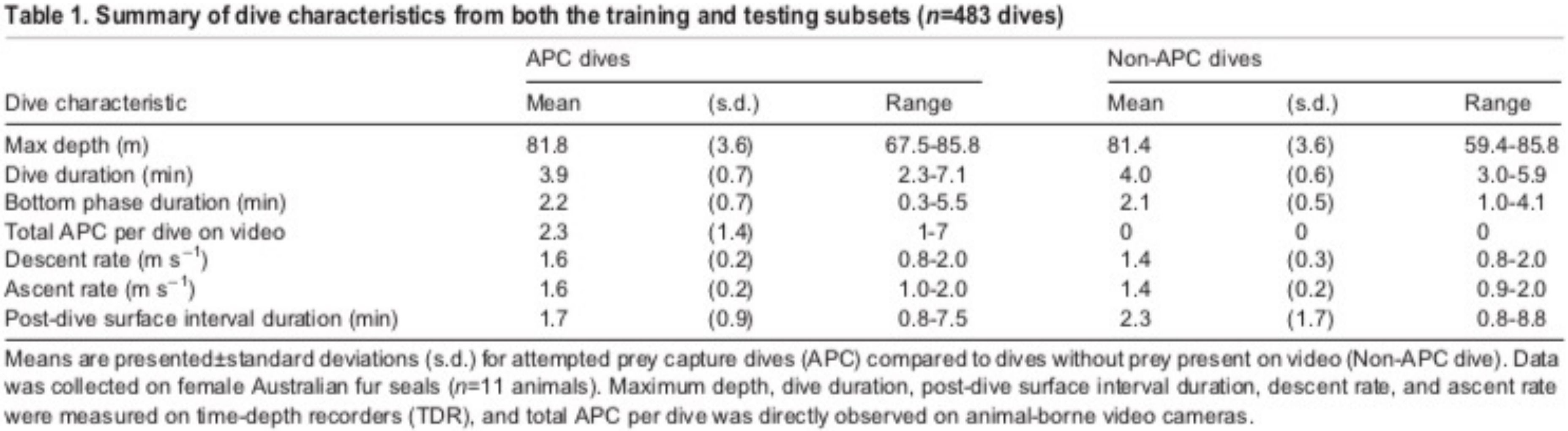


Table 1. Summary of dive characteristics from both the training and testing subsets (*n*=483 dives)

	APC dives			Non-APC dives		
Dive characteristic	Mean	(s.d.)	Range	Mean	(s.d.)	Range
Max depth (m)	81.8	(3.6)	67.5-85.8	81.4	(3.6)	59.4-85.8
Dive duration (min)	3.9	(0.7)	2.3-7.1	4.0	(0.6)	3.0-5.9
Bottom phase duration (min)	2.2	(0.7)	0.3-5.5	2.1	(0.5)	1.0-4.1
Total APC per dive on video	2.3	(1.4)	1-7	0	0	0
Descent rate (m s^{-1})	1.6	(0.2)	0.8-2.0	1.4	(0.3)	0.8-2.0
Ascent rate (m s^{-1})	1.6	(0.2)	1.0-2.0	1.4	(0.2)	0.9-2.0
Post-dive surface interval duration (min)	1.7	(0.9)	0.8-7.5	2.3	(1.7)	0.8-8.8

Means are presented±standard deviations (s.d.) for attempted prey capture dives (APC) compared to dives without prey present on video (Non-APC dive). Data was collected on female Australian fur seals (*n*=11 animals). Maximum depth, dive duration, post-dive surface interval duration, descent rate, and ascent rate were measured on time-depth recorders (TDR), and total APC per dive was directly observed on animal-borne video cameras.

Volpov et al 2016 Biology Open

Note that the “best” models are in **bold**, but all models are shown, so not focusing on only a single model

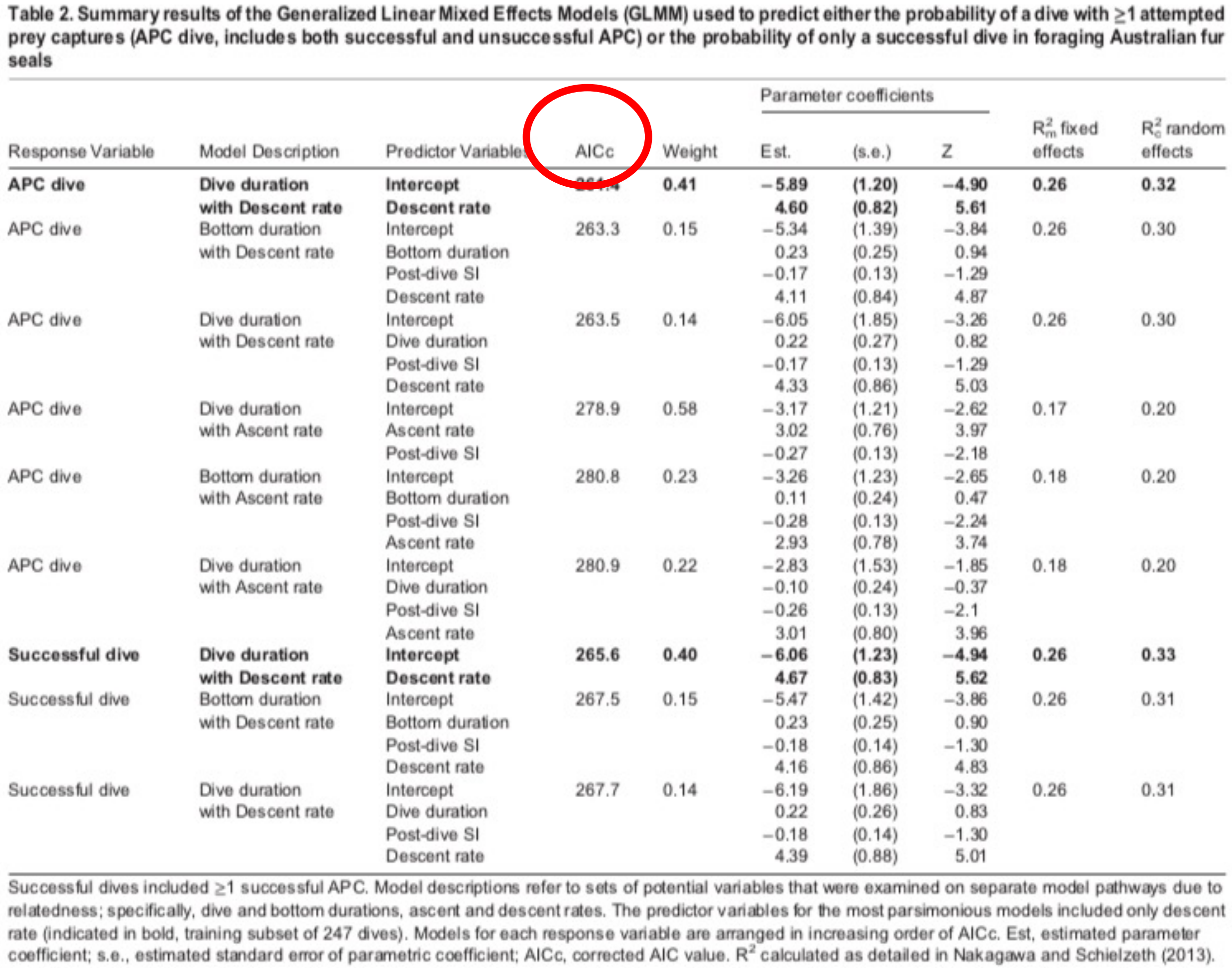


Table 2. Summary results of the Generalized Linear Mixed Effects Models (GLMM) used to predict either the probability of a dive with ≥1 attempted prey captures (APC dive, includes both successful and unsuccessful APC) or the probability of only a successful dive in foraging Australian fur seals

					Parameter coefficients				
Response Variable	Model Description	Predictor Variables	AICc	Weight	Est.	(s.e.)	Z	R^2_m fixed effects	R^2_c random effects
APC dive	**Dive duration with Descent rate**	**Intercept**	**261.4**	**0.41**	**−5.89**	**(1.20)**	**−4.90**	**0.26**	**0.32**
		Descent rate			**4.60**	**(0.82)**	**5.61**		
APC dive	Bottom duration with Descent rate	Intercept	263.3	0.15	−5.34	(1.39)	−3.84	0.26	0.30
		Bottom duration			0.23	(0.25)	0.94		
		Post-dive SI			−0.17	(0.13)	−1.29		
		Descent rate			4.11	(0.84)	4.87		
APC dive	Dive duration with Descent rate	Intercept	263.5	0.14	−6.05	(1.85)	−3.26	0.26	0.30
		Dive duration			0.22	(0.27)	0.82		
		Post-dive SI			−0.17	(0.13)	−1.29		
		Descent rate			4.33	(0.86)	5.03		
APC dive	Dive duration with Ascent rate	Intercept	278.9	0.58	−3.17	(1.21)	−2.62	0.17	0.20
		Ascent rate			3.02	(0.76)	3.97		
		Post-dive SI			−0.27	(0.13)	−2.18		
APC dive	Bottom duration with Ascent rate	Intercept	280.8	0.23	−3.26	(1.23)	−2.65	0.18	0.20
		Bottom duration			0.11	(0.24)	0.47		
		Post-dive SI			−0.28	(0.13)	−2.24		
		Ascent rate			2.93	(0.78)	3.74		
APC dive	Dive duration with Ascent rate	Intercept	280.9	0.22	−2.83	(1.53)	−1.85	0.18	0.20
		Dive duration			−0.10	(0.24)	−0.37		
		Post-dive SI			−0.26	(0.13)	−2.1		
		Ascent rate			3.01	(0.80)	3.96		
Successful dive	**Dive duration with Descent rate**	**Intercept**	**265.6**	**0.40**	**−6.06**	**(1.23)**	**−4.94**	**0.26**	**0.33**
		Descent rate			**4.67**	**(0.83)**	**5.62**		
Successful dive	Bottom duration with Descent rate	Intercept	267.5	0.15	−5.47	(1.42)	−3.86	0.26	0.31
		Bottom duration			0.23	(0.25)	0.90		
		Post-dive SI			−0.18	(0.14)	−1.30		
		Descent rate			4.16	(0.86)	4.83		
Successful dive	Dive duration with Descent rate	Intercept	267.7	0.14	−6.19	(1.86)	−3.32	0.26	0.31
		Dive duration			0.22	(0.26)	0.83		
		Post-dive SI			−0.18	(0.14)	−1.30		
		Descent rate			4.39	(0.88)	5.01		

Successful dives included ≥1 successful APC. Model descriptions refer to sets of potential variables that were examined on separate model pathways due to relatedness; specifically, dive and bottom durations, ascent and descent rates. The predictor variables for the most parsimonious models included only descent rate (indicated in bold, training subset of 247 dives). Models for each response variable are arranged in increasing order of AICc. Est, estimated parameter coefficient; s.e., estimated standard error of parametric coefficient; AICc, corrected AIC value. R^2 calculated as detailed in Nakagawa and Schielzeth (2013).

Used AICc for small sample size correction

Volpov et al 2016 Biology Open

Descent rate was the ‘best’ predictor of the probability of a dive that had prey present (but **all** models are presented in table)

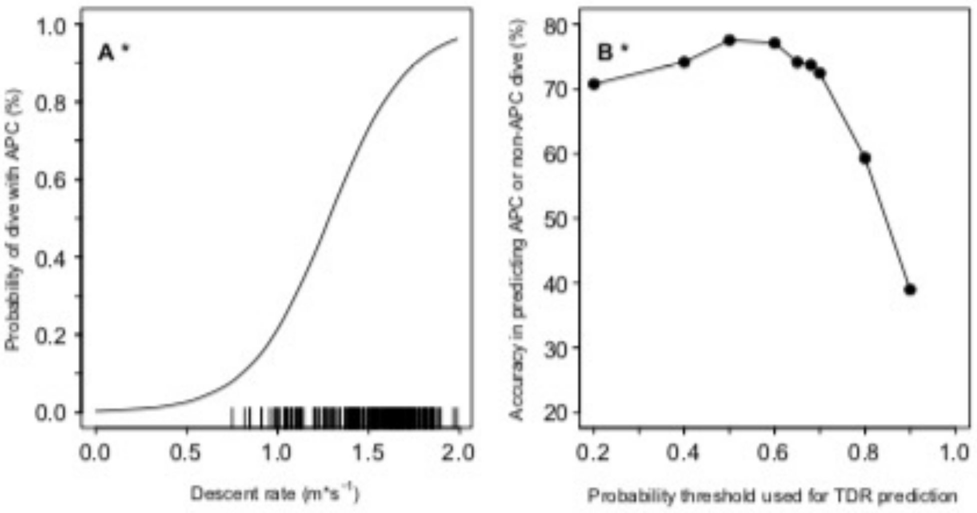


Fig. 1. Probability of a dive with ≥1 attempted prey captures (APC) in response to descent rate and accuracy of the GLMM relative to animal-borne video. (A) The most parsimonious model on the training subset included descent rate as predictive variable (Table 2). Distribution of descent rate is indicated with a rug plot. (B) Accuracy was calculated as the percent of dives correctly predicted as either APC or non-APC on the testing subset of dives (Table S1).

Volpov et al 2016 Biology Open

Cross-validation

Optional additional example on cross-validation

Cross-validation

- You can also do a variety of cross-validation experimental designs to **test predictions**
- **Training dataset:** build model on this data
- **Testing or validation dataset:** test model predictions on this new data
- **Split data into training and testing dataset**
    - Train model on 3 animals, and Hold-out 1 animal to test model
    - Randomly subset 50% of each animal's data points to put in test or train

Optional additional example on cross-validation

Example of Cross-validation and forwards model selection

Can head accelerometers predict prey capture success in foraging fur seals?

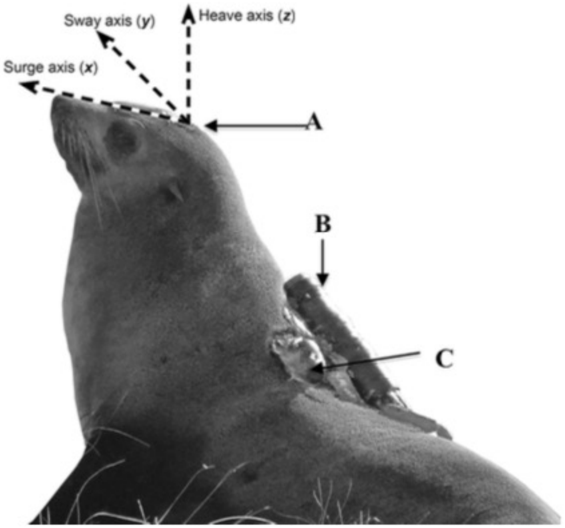


Fig 1. Photo of dataloggers deployed on Australian fur seals. Seals were instrumented with (A) accelerometer measuring surge (anterior-posterior), sway (lateral), and heave (dorsal-ventral) (B) National Geographic Crittercam measuring video, and (C) time-depth-recorder.

doi:10.1371/journal.pone.0128789.g001

This is a contingency table or confusion matrix below

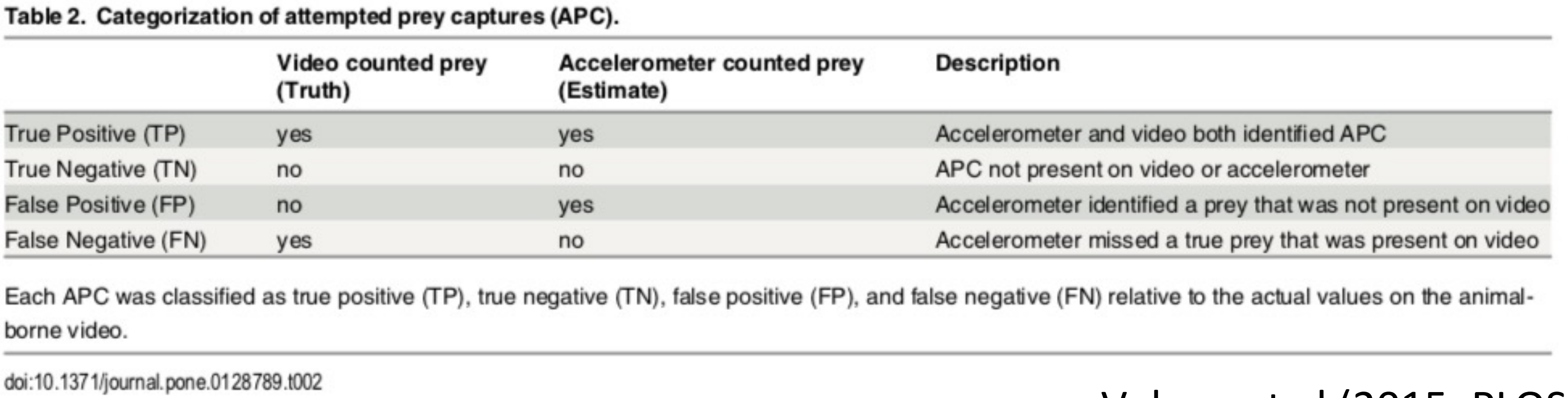


Table 2. Categorization of attempted prey captures (APC).

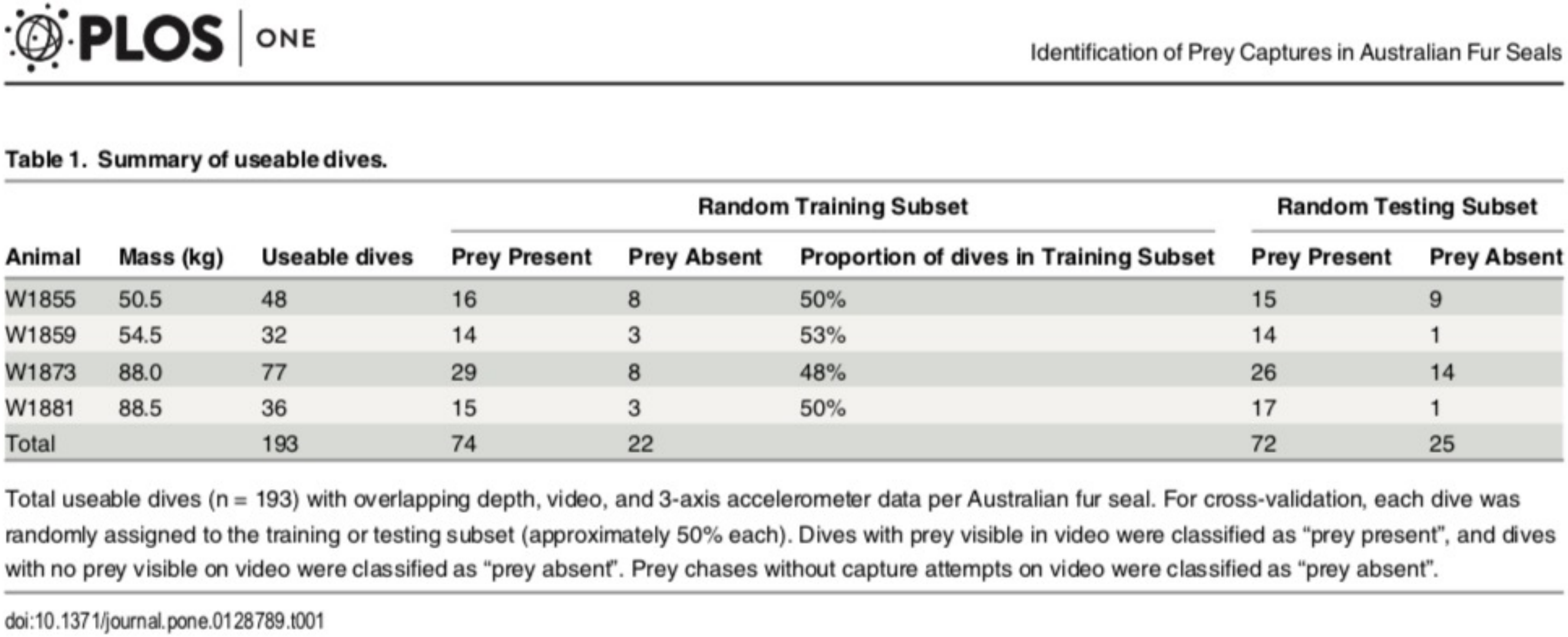
	Video counted prey (Truth)	Accelerometer counted prey (Estimate)	Description
True Positive (TP)	yes	yes	Accelerometer and video both identified APC
True Negative (TN)	no	no	APC not present on video or accelerometer
False Positive (FP)	no	yes	Accelerometer identified a prey that was not present on video
False Negative (FN)	yes	no	Accelerometer missed a true prey that was present on video

Each APC was classified as true positive (TP), true negative (TN), false positive (FP), and false negative (FN) relative to the actual values on the animal-borne video.

doi:10.1371/journal.pone.0128789.t002

Volpov et al (2015, PLOSONE)

- **LME model analysis with animal as random factor**
- **Split data into training and testing dataset**
  - Randomly subset 50% of each animal's data points to put in test or train
- **Training dataset:** build model on this data
- **Testing or validation dataset:** test model predictions on this new data



PLOS | ONE

Identification of Prey Captures in Australian Fur Seals

Table 1. Summary of useable dives.

			Random Training Subset			Random Testing Subset	
Animal	**Mass (kg)**	**Useable dives**	**Prey Present**	**Prey Absent**	**Proportion of dives in Training Subset**	**Prey Present**	**Prey Absent**
W1855	50.5	48	16	8	50%	15	9
W1859	54.5	32	14	3	53%	14	1
W1873	88.0	77	29	8	48%	26	14
W1881	88.5	36	15	3	50%	17	1
Total		193	74	22		72	25

Total useable dives (n = 193) with overlapping depth, video, and 3-axis accelerometer data per Australian fur seal. For cross-validation, each dive was randomly assigned to the training or testing subset (approximately 50% each). Dives with prey visible in video were classified as "prey present", and dives with no prey visible on video were classified as "prey absent". Prey chases without capture attempts on video were classified as "prey absent".

doi:10.1371/journal.pone.0128789.t001

Results of GLMM paper that us

- Some head movements recorded by the accelerometers were unrelated to prey
- 1 peak in acceleration did not always indicate 1 prey item.
- Accelerometers are a complementary tool for investigating foraging behaviour in pinnipeds, but that detection **and FP correction factors** need to be applied for reliable field application.

Pros and Cons of Cross-validation

- Pros
  - Assesses ability to predict on **new data**
- Cons
  - Takes more time in analysis and data collection
  - Reduces samples/animals that model is trained on
- Some scientists don’t really want to know their model has very low ability to predict on new data

What determines prediction errors?

- Prediction errors result from both bias and sampling variance (sampling error) in model parameter estimates. The effects of bias and sampling varianve are inversely related (the *bias-variance tradeoff*).
- The coefficients of the simplest model are likely to be biased, because the true relationship is likely to be more complex. But the coefficients of a simple model have relatively low sampling error (low sample variance) compared to a more complex model.
- The coefficients of complex models have lower bias (their long-run averages are close to their true values), but the coefficients of complex models have high sampling error (high sample variance).
- Prediction error is typically minimized somewhere in between.

Workshop

Workshop Thurs: Model Selection

- Scaling of basal metabolic rate (BMR) in mammals
  - Savage et al. (2004, Functional Ecology 18: 257-282)

- Bird abundance in forest fragments
  - Example with data dredging