Section 5.2.
Multiple linear regression

Highland Statistics Ltd.

highstat@highstat.com
www.highstat.com

Copyright Highland Statistics Ltd. www.highstat.com www.brodgar.com

RIKZ data

- Previous section:
  - Applied bivariate linear regression model
  - arbitrary used NAP as X

- More explanatory variables available:
  - grainsize,
  - humus,
  - angle of the beach,
  - exposure,
  - week,
  - etc.

Copyright Highland Statistics Ltd. www.highstat.com www.brodgar.com

RIKZ data

- In this section:
  - discuss **multiple** linear regression

- Allow one to model:
  - response variable (e.g. species richness)
  - as a linear function of multiple explanatory variables

- Hence the name:
  - multiple linear regression

Copyright Highland Statistics Ltd. www.highstat.com www.brodgar.com

Mathematical formulation multiple regression model:

$$Y_i = \alpha + \beta_1 X_{1i} + \ldots + \beta_p X_{pi} + \varepsilon_i$$

•Example RIKZ data:

$$R_i = \alpha \quad + \beta_1 NAP_i + \beta_2 Grainsize_i + \beta_3\, Humus_i + Week_i + \beta_4\, Angle_i + \varepsilon_i$$

Spot the difference

Copyright Highland Statistics Ltd. www.highstat.com www.brodgar.com

To reduce numerical output:

– concentrate on these 5 explanatory variables.

- Difference with bivariate regression:
  - Interpretation parameters

- The parameter β_1 shows:
  - change in species richness for a one-unit change in NAP
  - while holding all other variables constant.

Copyright Highland Statistics Ltd. www.highstat.com www.brodgar.com

- Same for other parameters.
- Partial regression slopes.
- Measure:
  - change in Y for particular X,
  - keeping remaining p-1 variables constant.

- What does that mean?

Copyright Highland Statistics Ltd. www.highstat.com www.brodgar.com

- Choose values for Grainsize, humus, angle, and pick a week:
  - Grainsize = 200
  - Humus = 50
  - Angle = 10
  - Week = 1

$$R_i = \alpha + \beta_1 NAP_i + \beta_2 \times 200 + \beta_3 \times 50 + \beta_4 \times 10 + 0 * 1 + \varepsilon_i$$

$$R_i = \text{constant} + \beta_1 NAP_i + \varepsilon_i$$

Copyright Highland Statistics Ltd. www.highstat.com www.brodgar.com

Source of variation	SS	Df	MS
Regression	$\sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2$	p	$\frac{\sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2}{p}$
Residual	$\sum_{i=1}^{n} (Y - \hat{Y}_i)^2$	n-p-1	$\frac{\sum_{i=1}^{n} (Y - \hat{Y}_i)^2}{n-p-1}$
Total	$\sum_{i=1}^{n} (Y_i - \bar{Y})^2$	n-1	

ANOVA table has similar form

Copyright Highland Statistics Ltd. www.highstat.com www.brodgar.com

- Null hypothesis:
  - All slopes are equal and 0
  - $H_0: \beta_1 = \beta_2 = .. = \beta_p = 0.$

- Just as in bivariate linear regression:
  - Ratio of
    - $MS_{regression}$
    - $MS_{residual}$
    - follows F distribution
- Ratio is used to test H_0.

- Here: F- statistic is 11.18
  - highly significant ($p < 0 .001$).

Copyright Highland Statistics Ltd. www.highstat.com www.brodgar.com

- Hence:
  - Hypothesis H_0: $\beta_1 = \ldots = \beta_5 = 0$ is rejected.

- F-statistic does not indicate whether one of the parameters is zero.

- t- test can be used for this.

Copyright Highland Statistics Ltd. www.highstat.com www.brodgar.com

- Estimated parameters, standard errors and t-values:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.30	7.97	1.17	0.25
angle2	0.02	0.04	0.39	0.70
NAP	-2.27	0.53	-4.30	<0.001
grainsize	0.00	0.02	0.11	0.92
humus	0.52	8.70	0.06	0.95
as.factor(week)2	-7.07	1.76	-4.01	<0.001
as.factor(week)3	-5.72	1.83	-3.13	<0.001
as.factor(week)4	-1.48	2.72	-0.55	0.59

Copyright Highland Statistics Ltd. www.highstat.com www.brodgar.com

What is the aim of building a model?

- Prediction?
  - Keep it as it is
- Know which covariates are important?
  - Drop the rubbish.
  - Easier to explain
  - Significant terms become more significant

- Week:
  - All levels go in, or week doesn't go in at all.
  - We need something that gives us one p-value for week
  - Put on the wish-list

Copyright Highland Statistics Ltd. www.highstat.com www.brodgar.com

- Here: know which covariates are important
- How to decide what to remove?
  - *Drop smallest beta?*
  - *All non-significant terms at once?*

- We need a protocol

- Collinearity spoils the fun
  - Increases SE, and therefore p-values
  - Will discuss VIFs later

Copyright Highland Statistics Ltd. www.highstat.com www.brodgar.com

IT approach
(Burnham and Anderson, 2002)
Specify a priori 10-15 models
Calculate differences in AIC
Model averaging

Hypothesis testing
t-test
F-test

Model Selection

Stepwise selection

Classical model selection approaches
AIC, CAIC or BIC

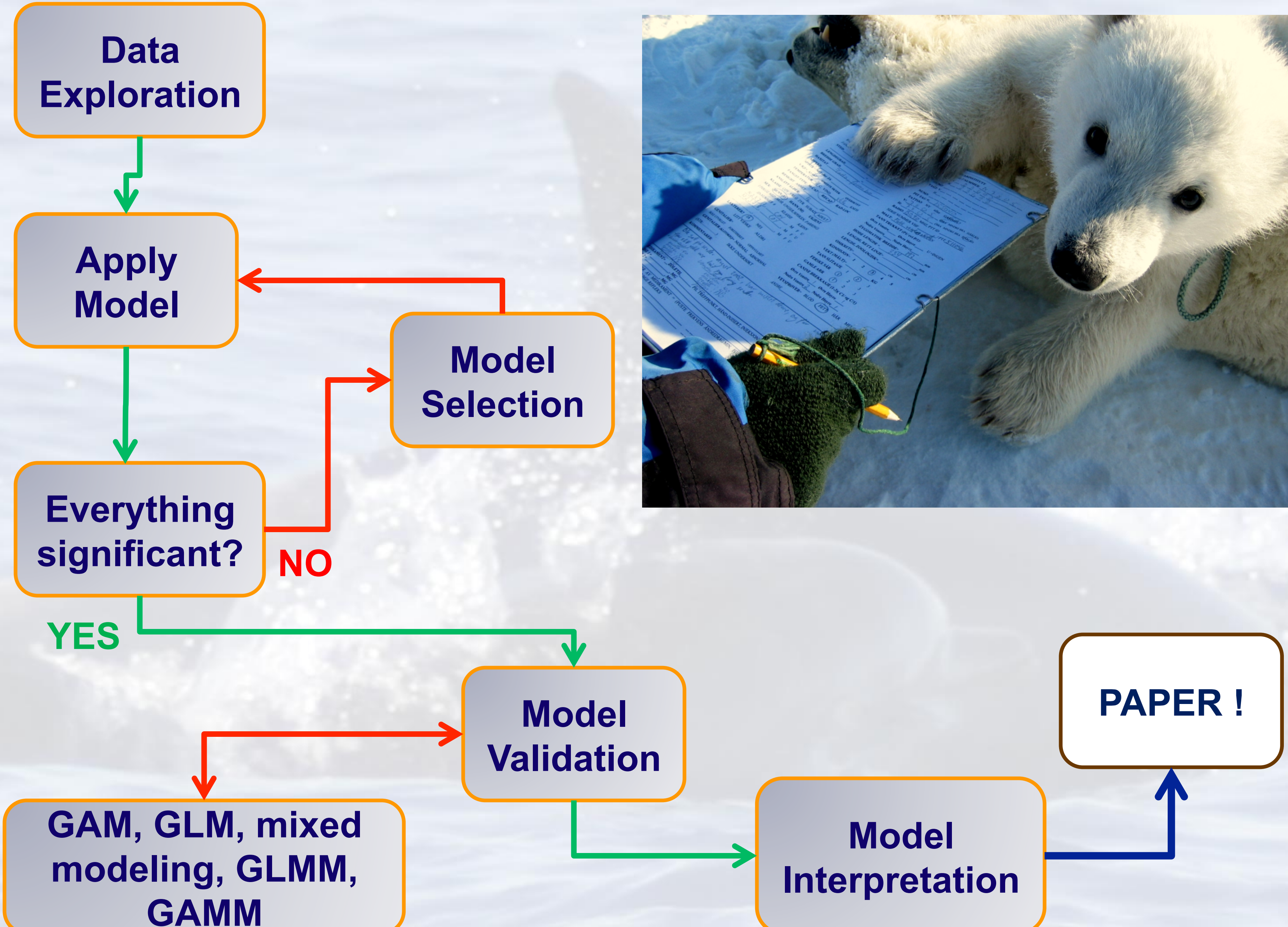
Do not do model selection at all!
Do it only on Interactions (Bolker, 2008)

Model selection

Four approaches:

1. Keep the model as it is
2. Hypothesis testing
   - t-statistics
   - F statistic
3. Use measure of goodness of fit
   - **AIC** or CAIC. BIC, Adjusted R^2
4. Information criteria
   - Burnham and Anderson (2002)

Copyright Highland Statistics Ltd. www.highstat.com www.brodgar.com


Data Exploration
Apply Model
Model Selection
Everything significant?
NO
YES
Model Validation
GAM, GLM, mixed modeling, GLMM, GAMM
Model Interpretation
PAPER !

1.Keep the model as it is

Or do only model selection on interaction

- Interactions are difficult to explain
- Bolker, 2008
- Mind **collinearity**!

Copyright Highland Statistics Ltd. www.highstat.com www.brodgar.com

2. Hypothesis testing approach

- Approach 1: *t*-statistic
  - Choose covariate with highest p-value
  - Drop and refit the model
  - Works OK for Gaussian distribution, and if there are no factors with >2 levels

Copyright Highland Statistics Ltd. www.highstat.com www.brodgar.com

Estimated parameters, standard errors and t-values:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.30	7.97	1.17	0.25
angle2	0.02	0.04	0.39	0.70
NAP	-2.27	0.53	-4.30	<0.001
grainsize	0.00	0.02	0.11	0.92
humus	0.52	8.70	0.06	0.95
as.factor(week)2	-7.07	1.76	-4.01	<0.001
as.factor(week)3	-5.72	1.83	-3.13	<0.001
as.factor(week)4	-1.48	2.72	-0.55	0.59

Drop

?

Copyright Highland Statistics Ltd. www.highstat.com www.brodgar.com

2. Hypothesis testing approach

Approach 2: F-statistic

- Compare two nested models.
- Definition of nested:
  - Two models are nested if we start with the full model and set some parameters equal to, for example, 0

Nested model: $Y_i = \alpha + \beta_1 \text{*Angle2} + \varepsilon_i$

Full model: $Y_i = \alpha + \beta_1 \text{*Angle2} + \beta_2 \text{*NAP} + \varepsilon_i$

Copyright Highland Statistics Ltd. www.highstat.com www.brodgar.com

Model 1: $Y_i = \alpha + \beta_1 * \text{Angle2} + \varepsilon_i$
Model 2: $Y_i = \alpha + \beta_1 * \text{Angle2} + \beta_2 * \text{NAP} + \varepsilon_i$

- Which model is always equal or better?
- Suppose H_0: $\beta_2 = 0$

- Calculate F statistic:

$$F = \frac{(RSS_1 - RSS_2)/(p-q)}{RSS_2/(n-p)}$$

- Big F-value is evidence against H_0

Copyright Highland Statistics Ltd. www.highstat.com www.brodgar.com

- For continuous variables:
  - F and t statistics give same p-values

Model 1: $R_i = \alpha + \beta_1 \, NAP_i + \beta_2 \, Grainsize_i + \beta_3 \, Humus_i + \beta_4 \, Angle_i + \varepsilon_i$

Model 2: $R_i = \alpha + \beta_1 \, NAP_i + \beta_2 \, Grainsize_i + \beta_3 \, Humus_i + \mathbf{Week_i} + \beta_4 \, Angle_i + \varepsilon_i$

Why do it?

Copyright Highland Statistics Ltd. www.highstat.com www.brodgar.com

What are we testing?

$$H_0 = \beta_{w2} = \beta_{w3} = \beta_{w4} = 0$$
$$H_a = \beta_{w2} = \beta_{w3} = \beta_{w4} \neq 0$$

F value = 6.19 *p-value* < 0.001

?

F-statistic gives one p-value for Week!

Copyright Highland Statistics Ltd. www.highstat.com www.brodgar.com

How to do it…

a) “Drop 1” R function

- Drop one explanatory variable
- Apply an F-test.
- Drop1 function again

	Df	Sum of Sq	RSS	AIC	F	Pr(F)
<none>		353.66	108.78			
angle2	1	1.46	355.12	106.96	0.15	0.70
NAP	1	176.37	530.03	124.98	18.45	0.00
grainsize	1	0.11	353.77	106.79	0.011	0.92
humus	1	0.03	353.70	106.78	0.004	0.95
as.factor(week)	3	177.51	531.17	121.08	6.190	0.00

Copyright Highland Statistics Ltd. www.highstat.com www.brodgar.com

How to do it…

b) Do it manually

Convince yourself that the drop1 is doing this:

Model 1: $R_i = \alpha + \beta_1 \text{ NAP}_i + \beta_2 \text{ Grainsize}_i + \beta_3 \text{ Humus}_i + \beta_4 \text{ Angle}_i + \varepsilon_i$

Model 2: $R_i = \alpha + \beta_1 \text{ NAP}_i + \beta_2 \text{ Grainsize}_i + \beta_3 \text{ Humus}_i + \mathbf{Week_i} + \beta_4 \text{ Angle}_i + \varepsilon_i$

Copyright Highland Statistics Ltd. www.highstat.com www.brodgar.com

anova (M1,M2)

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	37	353.66				
2	40	531.17	-3	-177.51	6.1902	0.00162

Same results!!

Copyright Highland Statistics Ltd. www.highstat.com www.brodgar.com

2. Hypothesis testing approach

Approach 3: "sequential F-statistic":

anova (M1)

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
angle2	1	124.86	124.86	13.06	0.001
NAP	1	319.32	319.32	33.41	<0.001
grainsize	1	106.76	106.76	11.17	0.002
humus	1	19.53	19.53	2.04	0.161
as.factor(week)	3	177.51	59.17	6.19	0.003
Residuals	37	353.66	9.56		

Copyright Highland Statistics Ltd. www.highstat.com www.brodgar.com

Spot the difference!

anova (M1,M2)

- ✓ **nested models**
- ✓ **drop 1 function in R**
- ✓ **trustable**

anova (M1)

- ✓ **sequential testing**
- ✓ **the order matters**
- ✓ **do not use it unless covariates are 100% independent**

Copyright Highland Statistics Ltd. www.highstat.com www.brodgar.com

3. Goodness of fit approach

- Use statistical criteria to select a model:
  - AIC, BIC, adjusted R^2, etc.
- Defined by:

$$AIC = n \log(SS_{residual}) + 2(p+1)$$

= how good is model + model complexity

Mind: Missing values / non-nested models

Various different definitions

Copyright Highland Statistics Ltd. www.highstat.com www.brodgar.com

- Lowest AIC is best model
- Highest is adjusted R^2 is best model

- Problem:
  - many possible combinations of covariates
  - Apply a backwards
  - or forwards selection
  - backwards is prefered

Copyright Highland Statistics Ltd. www.highstat.com www.brodgar.com

Demonstrate for:

$$R_i = \alpha + \beta_1 \, NAP_i + \beta_2 \, Grainsize_i + \beta_3 \, Humus_i + Week_i + \beta_4 \, Angle_i + \varepsilon_i$$

- See R code

Copyright Highland Statistics Ltd. www.highstat.com www.brodgar.com

Alternative

Adjusted r^2 = $$1-\frac{SS_{residual}/(n-(p+1))}{SS_{total}/(n-1)}$$

Is like R^2, but take into account n and p.

Copyright Highland Statistics Ltd. www.highstat.com www.brodgar.com

4. Information criteria

- IT approach

  (Burnhan and Anderson, 2000)
- Specify a priori 10-15 models
- Calculate differences in AIC
- Model averaging

We will discuss this in another exercise

Copyright Highland Statistics Ltd. www.highstat.com www.brodgar.com

Problems with selection procedures:

- Collinearity.
- Multiple comparisons

Collinearity:

- forward selection and backward selection might give different results.
- Avoid using covariates that represent same ecological signal!!

Copyright Highland Statistics Ltd. www.highstat.com www.brodgar.com

Problem multiple comparisons:

- Each time 5% chance making wrong statement.
- Apply large number of forward/backward selections:
  - this chance increases.

Copyright Highland Statistics Ltd. www.highstat.com www.brodgar.com

- Three ways to deal with this:
  - ignore the problem,
  - avoid using selection methods
  - apply a correction method
    - Bonferonni method.
- Bonferonni:
  - p-values are adjusted for the number of tests that are carried out.

Copyright Highland Statistics Ltd. www.highstat.com www.brodgar.com

- F test is for nested models
- AIC is for nested and non-nested models

Copyright Highland Statistics Ltd. www.highstat.com www.brodgar.com