



# Day3: Intro to linear models

## Model fitting

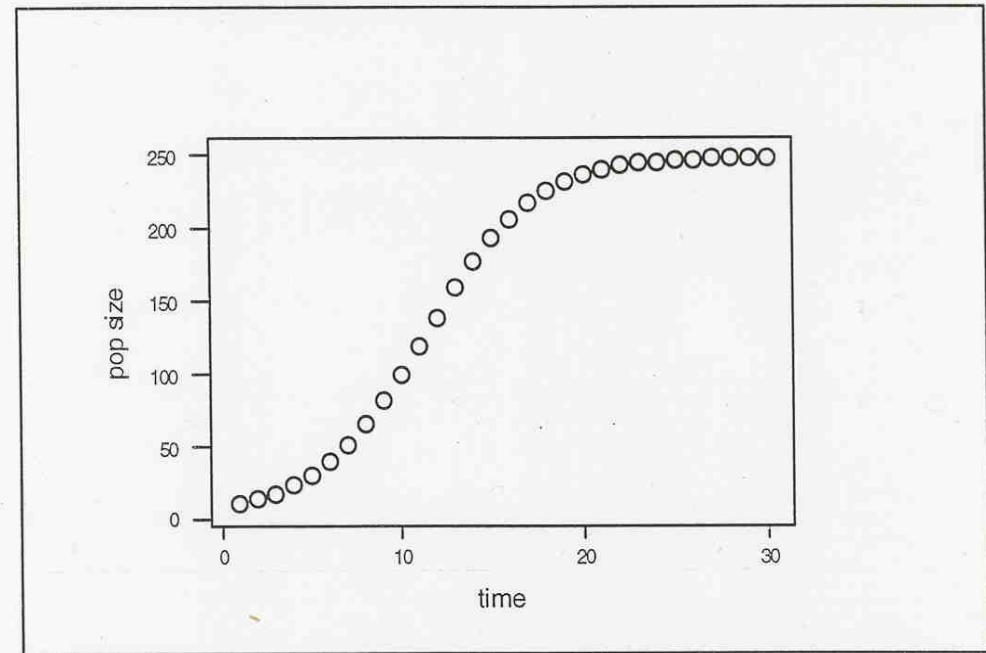


© Scott Adams, Inc./Dist. by UFS, Inc.

# What is a *model*?

- Models are used throughout science to simplify and explain the natural world
- e.g. logistic growth model simplifies population dynamics to 2 parameters:  $r$  and  $K$
  - **Statistical models** use data to reach conclusions...

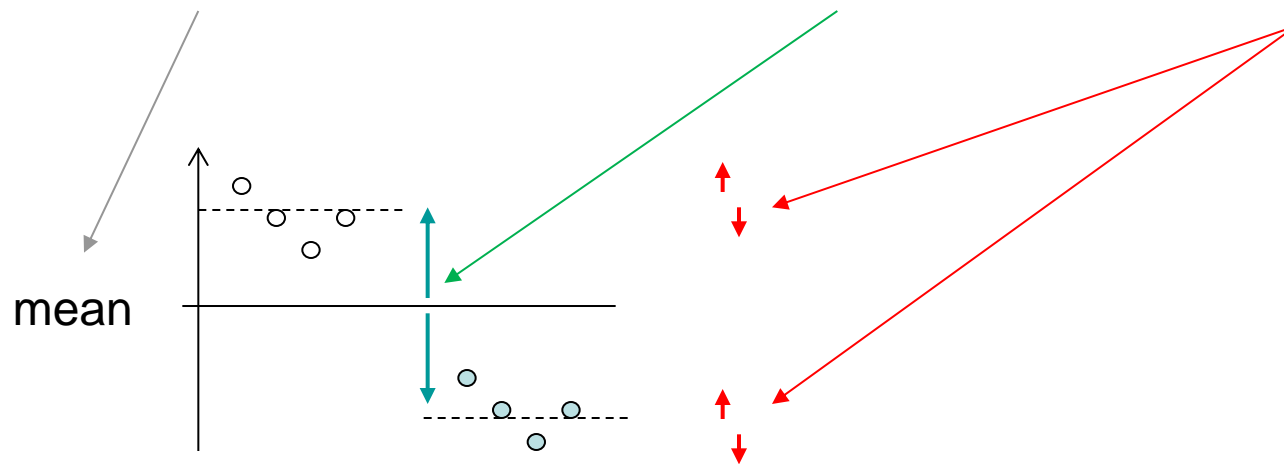
$$\frac{dN}{dt} = rN \left( \frac{K - N}{K} \right)$$



# Linear models

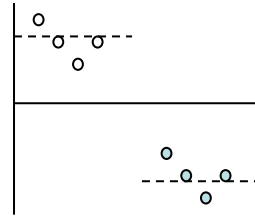
- A LM can model data (response variable) using a predictor variable(s).
  - Describing trends in the data by a simple equation
- For comparing treatments, the simple model is:

$$\text{Data} = \text{Mean} + \text{Treatment effect} + \text{Error}$$

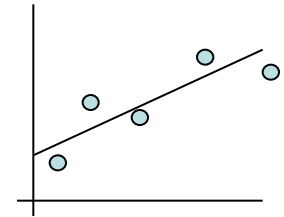


$$\text{Data} = \text{Mean} + \text{Treatment effect} + \text{Error}$$

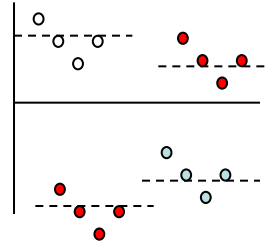
$$y = a + bx$$



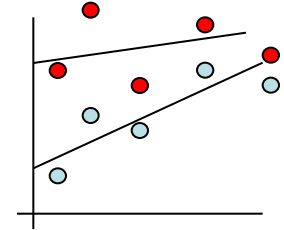
or



$$y = a + bx_1 + cx_2 + dx_1x_2$$



or



- It is a *linear* model because it is *additive*.
- Any model that has the *coefficients* (the  $a$ ,  $b$ ,  $c$ , etc) separated by  $+$  or  $-$  signs is linear.

$$\text{Data} = \text{Mean} + \text{Treatment effect} + \text{Error}$$

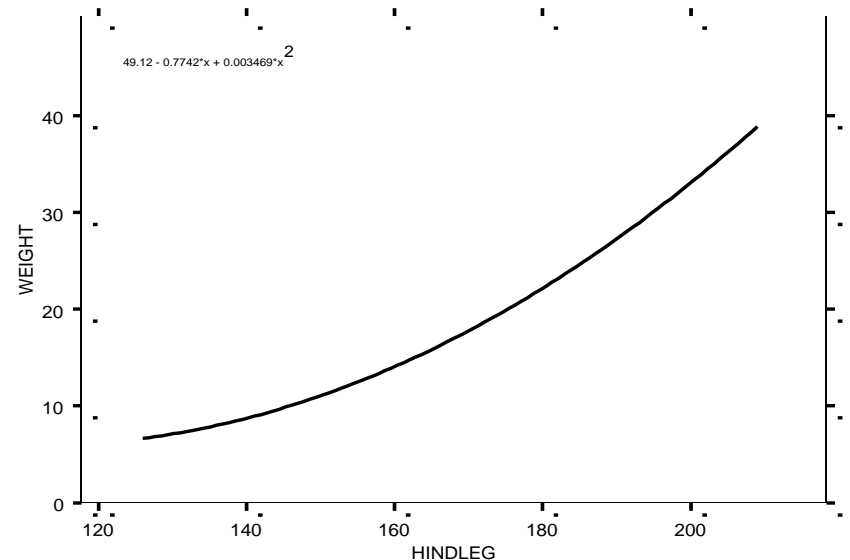
# “Non-linear”

- Linear models can sometimes describe non-linear patterns.
- This is a linear model but models a curve:

$$Y = a + bX + cX^2$$

- This is a non-linear model which also gives a curve:

$$Y = a + \frac{b}{c + X}$$



Non-linear because one predictor variable  $x$  is associated with two coefficients;  
cf model  $y = a + (b/c)x$  which is linear because the  $(b/c)$  is a constant ratio & could be a single coefficient (e.g.  $d = b/c$ )

# (General) Linear Models in R

Data = Mean + Treatment effect + 1

Sex	ht	sex x	grand x
M	1.82	1.808	1.75
M	1.81	1.808	1.75
M	1.79	1.808	1.75
M	1.80	1.808	1.75
M	1.82	1.808	1.75
F	1.68	1.692	1.75
F	1.69	1.692	1.75
F	1.70	1.692	1.75
F	1.70	1.692	1.75
F	1.69	1.692	1.75

> anova(model.glm)

General Linear Model: height versus sex

Factor	Type	Levels	Values
sex	fixed	2	1 2

Analysis of Variance for height, using Adjusted SS for Tests

Source	DF	Seq SS	Adj SS	Adj MS	F
sex	1	0.033640	0.033640	0.033640	280.33
Error	8	0.000960	0.000960	0.000120	
Total	9	0.034600			

$$\sum_{h=1}^{10} (ht - \bar{x})^2$$

$$\sum_{k=1}^2 (\bar{x}_k - \bar{x})^2$$

$$\sum_{h=1}^{10} (ht - \bar{x}_k)^2$$

$$\text{Data} = \text{Mean} + \text{Treatment effect} + \text{Error}$$

# Predictions are Coefficients

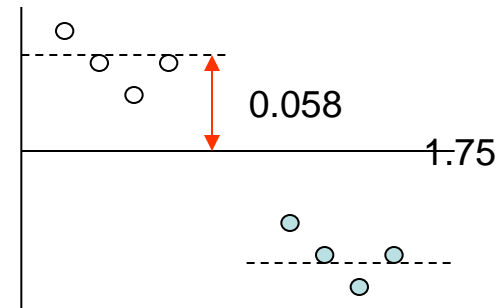
- Predicted values are calculated from the model by just ignoring the error:

$$\text{Height} = 1.75 \pm 0.058 \text{ (sex effect)}$$

- male height = 1.75m + 5.8cm = 1.808m
- female height = 1.75m - 5.8cm = 1.692m

```
> summary(model.glm)
```

Term	Coef	SE Coef	T	P
intercept	1.75000	0.00346	505.18	0.000
sex1	0.058000	0.003464	16.74	0.000





$$\text{Data} = \text{Mean} + \text{Treatment effect} + \text{Error}$$

## Predictions - concerns

- If we transformed the response, we need to think about back-transformations:
  - $\ln(Y) = a + b.\ln(X) + \varepsilon \rightarrow Y = \exp(a).X^b.\exp(\varepsilon)$   
so if  $X$  doubles,  $Y$  multiplies by  $2^b$
  - Work out upper & lower 95% CIs for  $Y$ , then back-transform these
- If we transformed any predictors, there's more, e.g.:
  - $Y = a + b.X^2 + \varepsilon \rightarrow$  if  $X$  doubles,  $Y$  goes up 4-fold
  - $Y = a + b.\ln(X) + \varepsilon \rightarrow$  if  $X$  doubles,  $Y$  adds on  $b.\ln(2)$
- To handle factor coefficients correctly, we need to understand...

$$\text{Data} = \text{Mean} + \text{Treatment effect} + \text{Error}$$

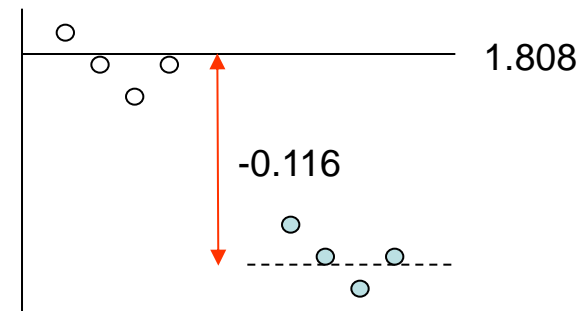
# Contrasts – different methods in R

```
> options("contrasts")
```

## Contrast “treatment”

Term	Coef	SE	Coef
Constant	1.80800	0.00346	
sex2	-0.11600	0.003464	

This is the default in R



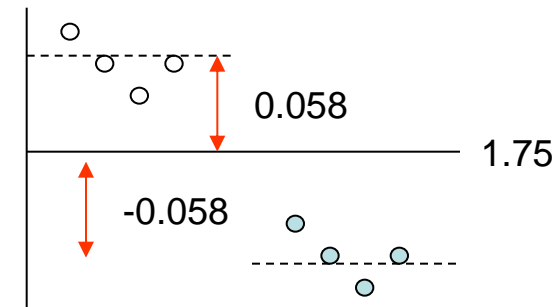
The contrast is with a particular *treatment* group (e.g. control)

```
> options( contrasts= c(contr.sum,contr.poly) )
```

## Contrast “sum”

Term	Coef	SE	Coef
Constant	1.75000	0.00346	
sex1	0.058000	0.003464	

Find the other coefficient by subtracting the one(s) given from 0



Coefs *sum* to zero within a group

$$\text{Data} = \text{Mean} + \text{Treatment effect} + \text{Error}$$

# Testing Null hypotheses

General Linear Model: height versus sex

Factor	Type	Levels	Values
sex	fixed	2	1 2

Analysis of Variance for height, using Adjusted SS for Tests

Source	DF	Seq SS	Adj SS	Adj MS	F	P
sex	1	0.033640	0.033640	0.033640	280.33	0.000
Error	8	0.000960	0.000960	0.000120		
Total	9	0.034600				

**$P < 0.0005$**

# (General) Linear Models

LM description	Traditional method
Predictor is a 2-level factor	= Two-sample t-test
Predictor is a 2-level factor, crossed with error blocks	= Paired t-test
Predictor is a 3+ -level factor	= One-way ANOVA
2 predictors, both factors	= Two-way ANOVA
2+ predictors, nested	= Nested ANOVA
Predictor is a continuous variable	= Regression
2+ continuous predictors	= Multiple regression
Mix of continuous and discrete predictors	= ANCOVA
Some predictor/s specify non-treatment groups	= Model 2 ANOVA

# (General) Linear Models

- Simple linear model

$$y = a + bx + \varepsilon$$

e.g.  $\text{WEIGHT} = 1.27 + 0.30 * \text{HLEG}$

- Simple linear model with transformation

$$y = a + b \cdot \log(x) + \varepsilon$$

e.g.  $\text{WEIGHT} = 0.23 + 0.01 * \log(\text{HLEG})$

- Polynomial regression (e.g. quadratic)

$$y = a + bx + cx^2 + \dots + mx^n + \varepsilon$$

e.g.  $\text{WEIGHT} = 0.11 + 0.27 * \text{HLEG} + 0.13 * \text{HLEG}^2$

- Multiple regression

$$y = a + bx_1 + cx_2 + \dots + mx_n + \varepsilon$$

e.g.  $\text{WEIGHT} = 0.25 + 0.89 * \text{HLEG} + 0.46 * \text{FLEG}$

- Multiple regression with interactions

$$y = a + bx_1 + cx_2 + dx_1x_2 + \dots + \varepsilon$$

etc.



# (General) Linear Models

In R....

```
> lm(y ~ x)
```

```
> lm(y ~ log(x))
```

```
> lm(y ~ x + I(x^2))
```

```
  or lm(y ~ poly(x, 2))
```

```
> lm(y ~ x1 + x2 + x3)
```

```
> lm(y ~ x1 * x2 * x3)
```

```
  or lm(y ~ (x1 + x2 + x3)^3)
```

# LM, GLM

- The term “linear model” either refers to model structure ( $Y=a+bX$ ) or is used to mean a general linear model.
  - A (general) LM is any LM that assumes residuals follow a Normal (Gaussian) distribution
- A Generalised Linear Model (GLM) is a model where error distributions need NOT be Normal (*tomorrow*)

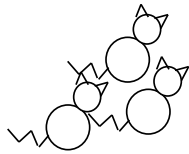


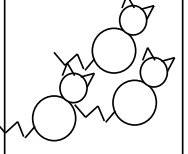
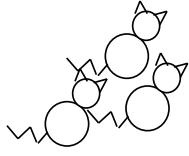

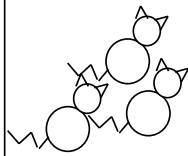
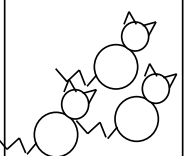
# Fitting LMs

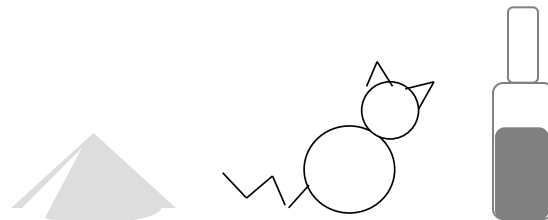


# Background: Factorial designs and interactions

# E is for ...Experiment

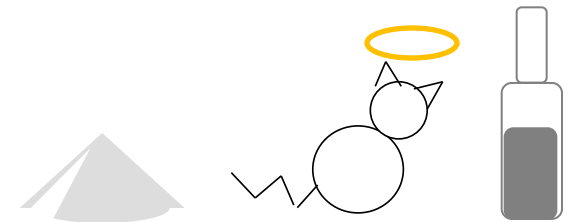
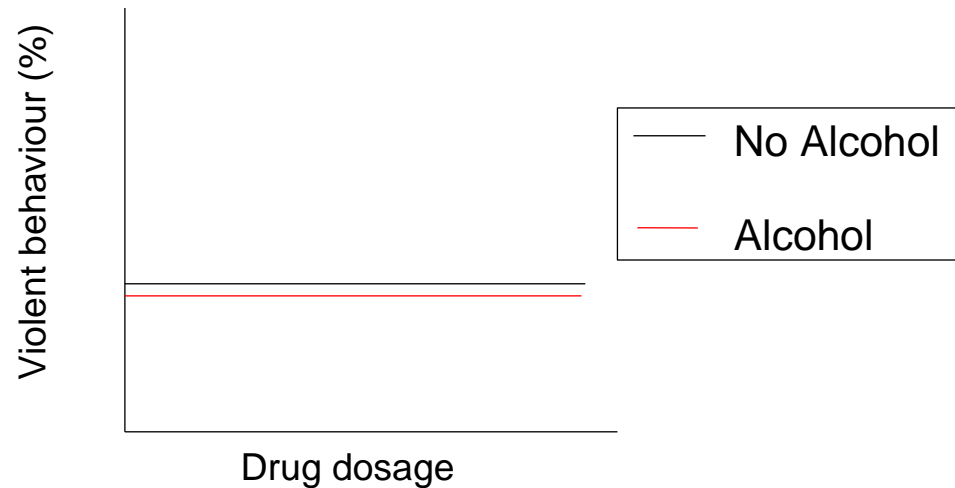
- We are interested in the effects of a new drug on the prevalence of violent behaviour in mice.
- As drugs are often taken with alcohol – especially by mice – we are also interested in the **interaction** between the drug and alcohol.
- So we design a **full-factorial** experiment.

Dose (mg) ->	0	25	50	75
Alcohol				
No Alcohol				

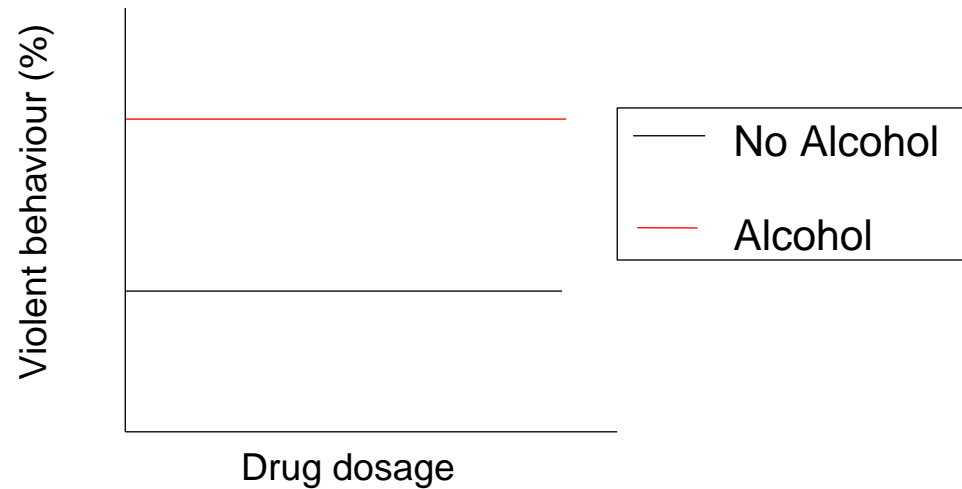


*There are 5 simple  
conclusions we might get  
from our experiment...*

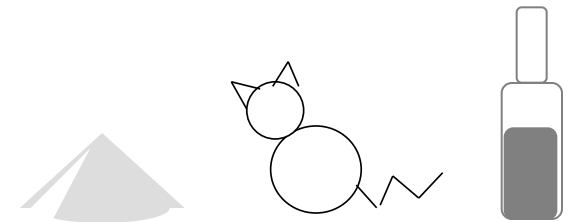
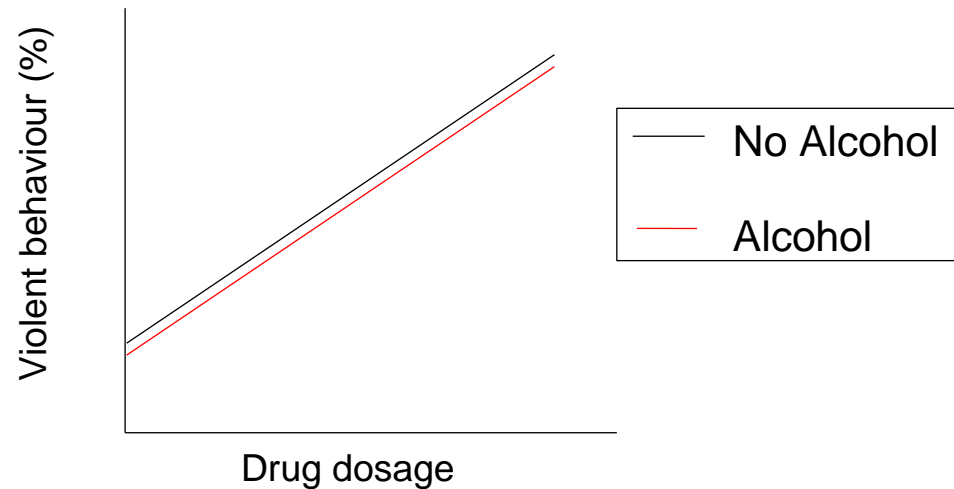
# 1. Neither the drug dosage, nor alcohol, affects behaviour



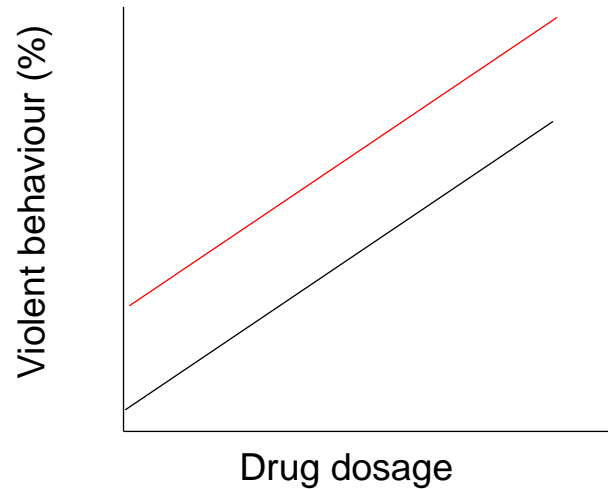
## 2. Alcohol affects behaviour, drug dosage doesn't



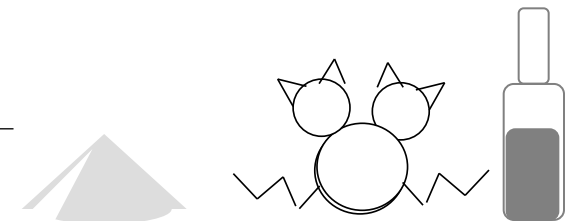
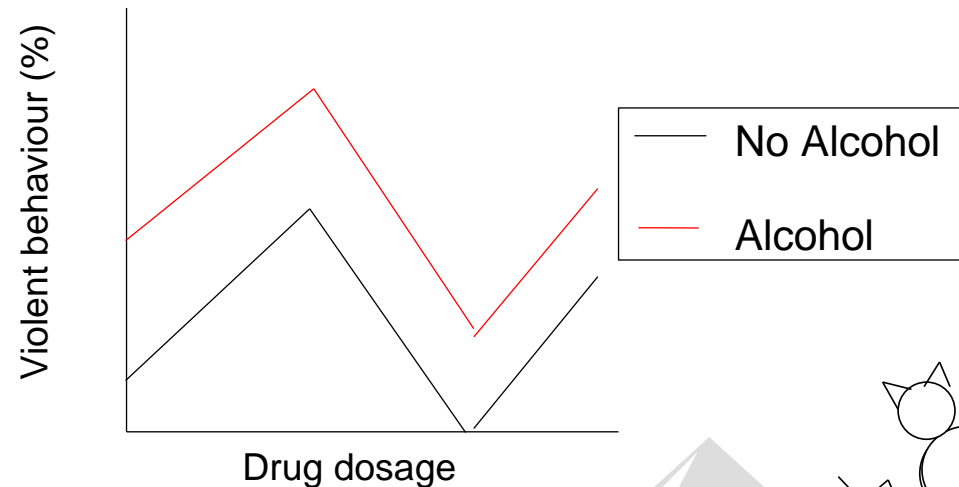
### 3. Drug dosage affects behaviour, alcohol doesn't



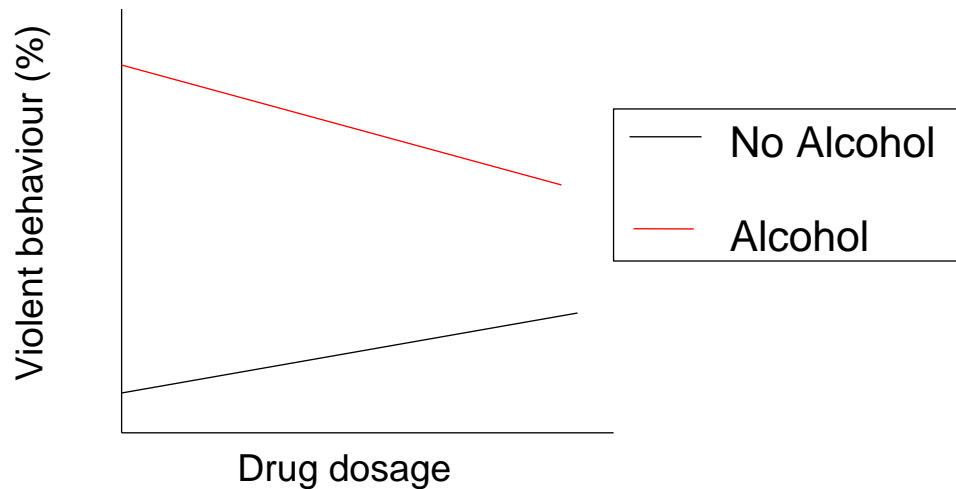
# 4. Both alcohol and drug dosage affect behaviour



*Additive effects:  
whatever the drug  
dosage, alcohol always  
affects behaviour in the  
same way*



## 5. Both alcohol and drug affect behaviour, and they interact



The lines are not parallel, i.e.

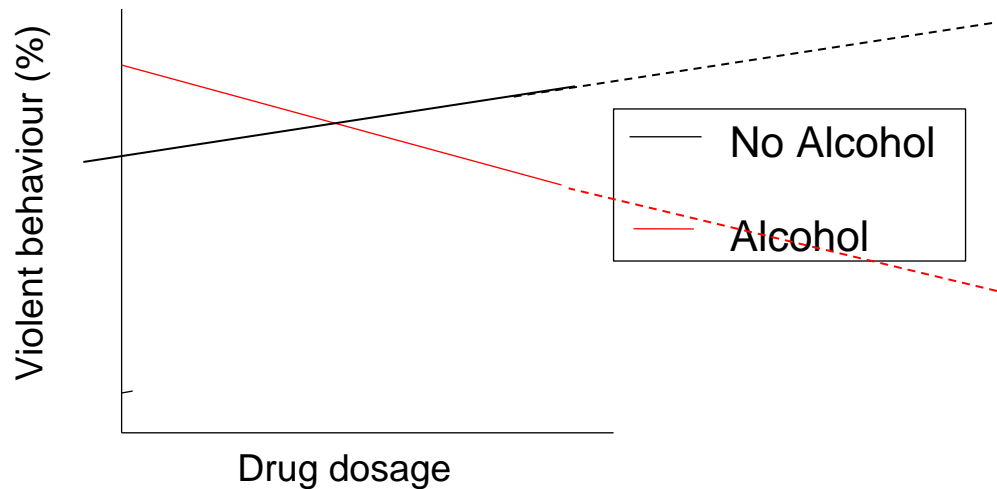
- *the effect of dosage depends on whether or not alcohol is involved and*
- *the effect of alcohol depends on the dosage.*

*Non-additive effects of alcohol and drug*





# Have we missed one?

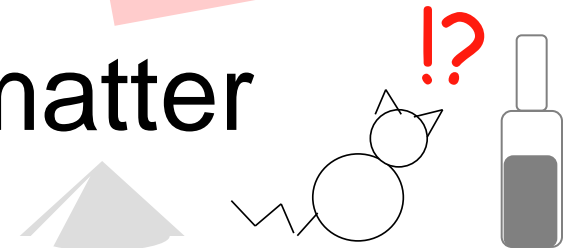


6. “Alcohol doesn’t affect behaviour, but alcohol and drug interact”

No!

Interaction => both factors matter

• “marginality”



# Practical

- Linear Models Part I and Part II

# Fundamental Objectives

- |  |  |  |
|--|--|--|
| <ul style="list-style-type: none"><li>• <u>Hypothesis Testing</u></li><li>• Did my treatments have the hypothesized effects?</li></ul> | <ul style="list-style-type: none"><li>• <u>Pattern Exploration &amp; understanding</u></li><li>• What combination of variables can be reliably related to the dependent variable?</li><li>• What's the best estimate for each coefficient?</li></ul> | <ul style="list-style-type: none"><li>• <u>Prediction</u></li><li>• What combination of variables provides most accurate prediction?</li></ul> |
|--|--|--|

Evaluating a  
planned design:  
Full model

Seeking a minimum  
adequate model;  
comparing models

Seeking a  
min(AIC) or  
max( $R^2$ )

# A good principle

- If you have a designed experiment or precise hypotheses, use the statistical model that corresponds to the design.
- Example: limpets.csv
- If you are exploring your dataset, simplify the model stepwise, justified by biological reasoning.
- Use model selection methods if you have alternative “candidate models” based on different variables.
- Example: Soay.csv

# Steps to a useful model

- Design the Experiment / Sampling / Observation
  - Know what your aim is (hypothesis?)
  - Randomisation, Replication
- Collect data
- Plot data
- Fit a model
- Examine diagnostics to check model assumptions
  - Residual plots
- Examine the model
  - ANOVA table?
  - Coefficients
  - Goodness of fit
- Simplify or compare models?

# Designed Experiment: 2-way factorial ANOVA



Limpet

Planned Experiment:

Egg numbers vs.

- Density
  - Density-dependence?
  - 5 levels (factor or quant.?)
- Seasonal differences
  - 2 levels
- 3 reps

Hypothesis



Null hypothesis: No effect of density or season  
Alternatives 1, 2: Effect of Season OR Density  
Alternative 3: Additive effects of Season AND Density  
Alternative 4: Multiplicative Effect (INTERACTION)

# Data exploration

- WEIGHT of sheep
  - response variable
- Could be a function of
  - Sex
  - Age
  - Parasite Load
  - etc.



Soay sheep

But if we fit a full model with all possible effects, coefficients will be poorly estimated – large SEs

# Procedure

- Not a planned experiment, so:
- start with full model;
- explore model and simplifications, to:
- find a minimum adequate model (MAM).

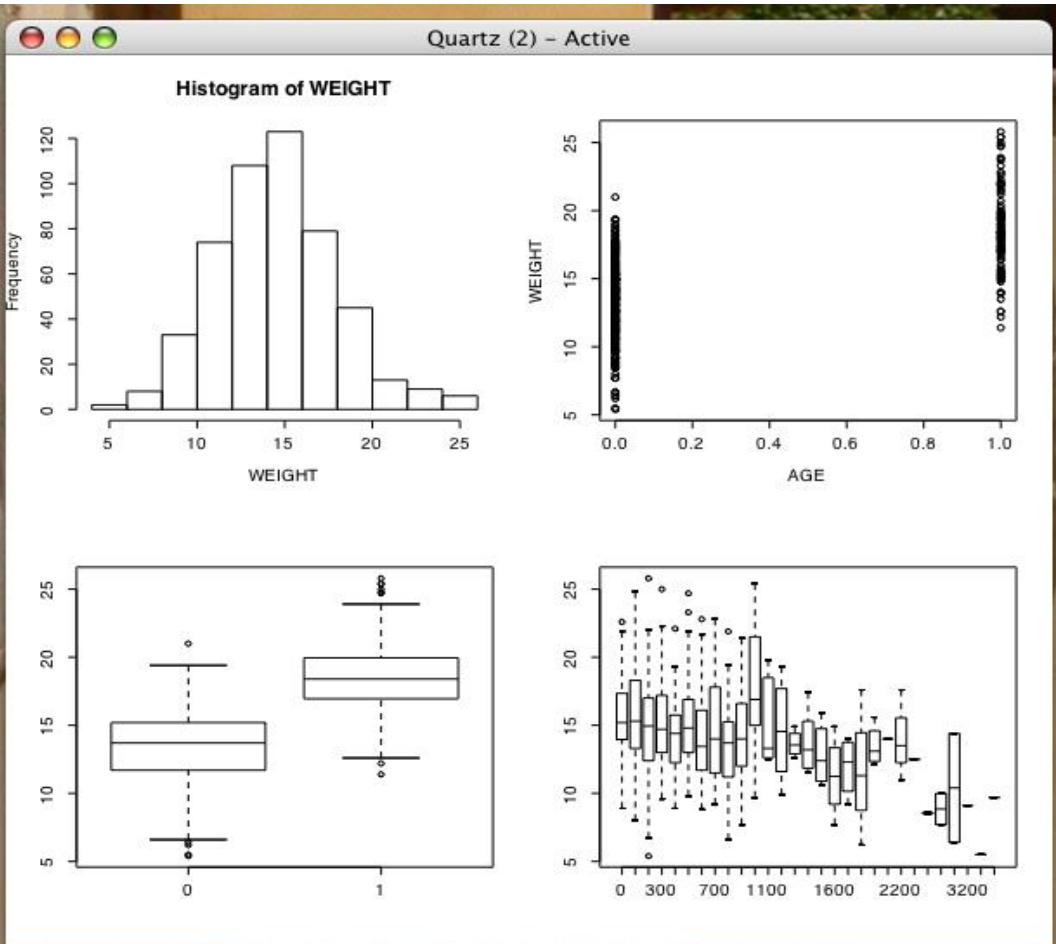
## Aims:

- Test null hypotheses about our predictors...
  - No effect of some predictors?
  - Predictors don't interact?
- Optimise the model so we get the best set of parameter estimates for future predictions?

- Full model:
  - one 3-way, three 2-way interactions; three main effects



# Plot the data



## Greetings Master - Loading Your Profile

Loading required package: lattice  
Note: The default device has been opened to honour attempt to modify trellis settings

Loading required package: graphics  
Loading required package: grDevices  
Loading required package: stats

Trellis Will be Pretty; MASS engaged!

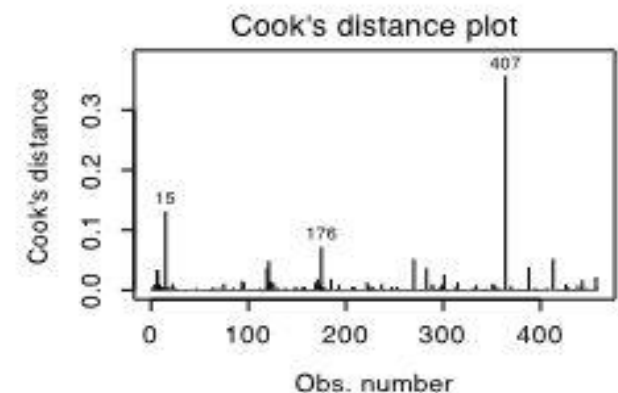
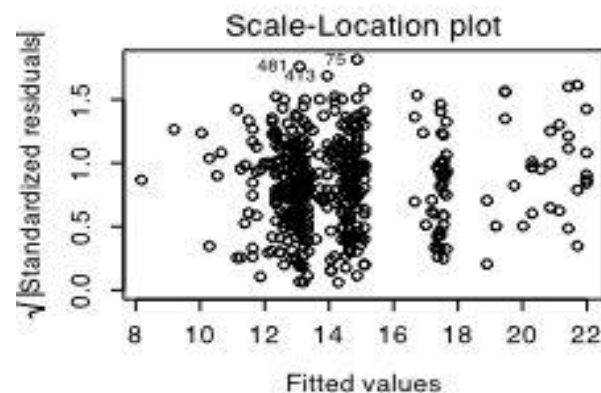
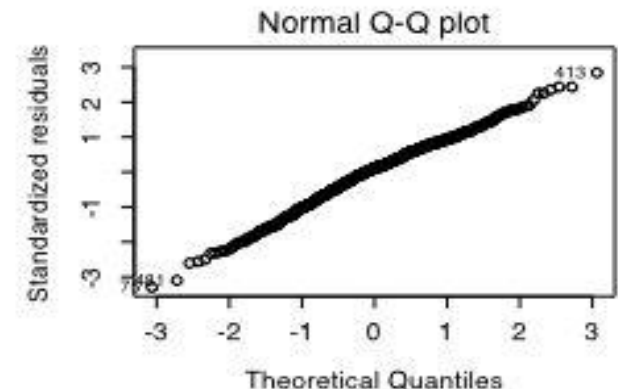
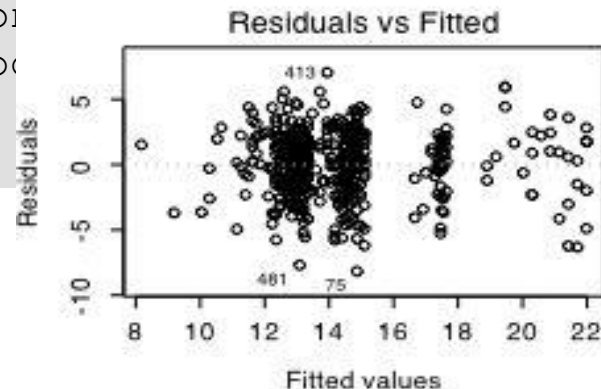
```
> soay<-read.csv("~/soay2.csv")
> names(soay)
[1] "NAME"      "YEARf"      "AGE"      "SEX"
[5] "WEIGHT"    "Testosterone" "ODIN"     "STR"
[9] "SURV1"
> par(mfrow=c(2,2))
> attach(soay)
> hist(WEIGHT)
> plot(WEIGHT~AGE)
> boxplot(WEIGHT~AGE)
> boxplot(WEIGHT~STR)
>
```

- **hist()**
- **plot()**
- **boxplot()**

# Fit the full model

## Examine the diagnostics

```
> m1 <- lm(W EIGHT ~fact or (AG E) * SEX*S TR) # our maximum model
> names(m1)
[1] "coef ficients" "res idu als" "efec ts" "rank "
[5] "fitted.values" "as"
[9] "na.action" "co"
[13] "terms" "mo"
> par (mfrow=c(2,2))
> plot(m1)
```



# Examine the model

ANOVA table with *sequential* SS: start from the bottom!

- Highest order term significant?

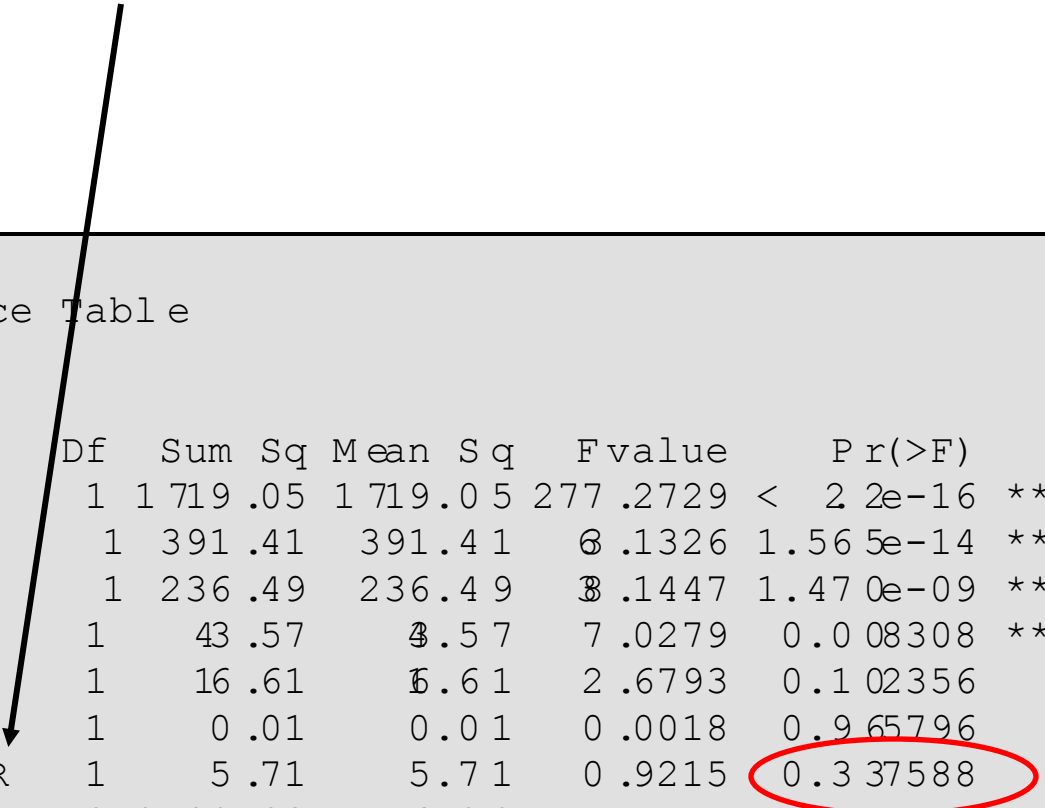
```
> anova(m1)
Analysis of Variance Table

Response: WEIGHT
```

	Df	Sum Sq	Mean Sq	Fvalue	Pr(>F)	
factor(AGE)	1	1719.05	1719.05	277.2729	< 2.2e-16	***
SEX	1	391.41	391.41	6.1326	1.565e-14	***
STR	1	236.49	236.49	3.1447	1.470e-09	***
factor(AGE):SEX	1	43.57	43.57	7.0279	0.008308	**
factor(AGE):STR	1	16.61	16.61	2.6793	0.102356	
SEX:STR	1	0.01	0.01	0.0018	0.965796	
factor(AGE):SEX:STR	1	5.71	5.71	0.9215	0.337588	
Residuals	450	2789.93	6.20			

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1



## Now: Simplify...

- Take the 3-way interaction out, and compare the model to the full model...
- Could the model without the 3-way interaction explain about as much variation as the model with it?
- (Does the Residual Sums of Squares change?)
- Do an  $F$ -test comparing the two residual SS...  

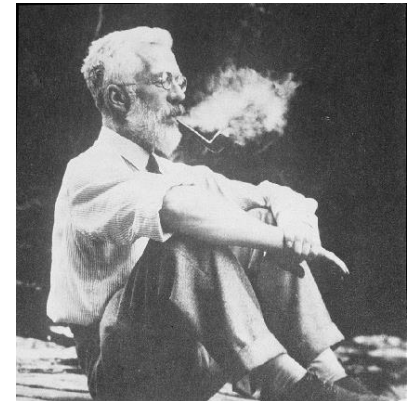
```
> anova(lm0, lm1)
```
- Unless  $P < 0.05$ , the term we removed may be superfluous.

# Single-Term Deletions

- The significance of a term in a model is the  $P$ -value for the comparison of the models with and without the term
- This is a Variance Ratio Test. You'll use a similar procedure to do Likelihood Ratio tests with GLMs
- Remember: if an interaction is significant, the main effects involved must be important too.

# Light at the end of the tunnel

- Eventually we'll get to a model where dropping any eligible terms makes the fit significantly worse.
- Sometimes...we should check robustness of our MAM by forwards and backwards routes of model simplification



# Practical

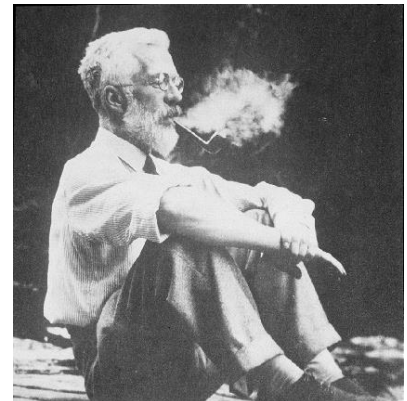
- Linear Models Part I and Part II

## Further reading

- Burnham & Anderson (2002): Model Selection and Multimodel Inference
- Crawley: The R Book
- Whittingham *et al*: J Appl Ecol 2006

# Light at the end of the tunnel

- Eventually we'll get to a model where dropping any eligible terms makes the fit significantly worse.
- We should check robustness of our MAM by forwards and backwards routes of model simplification
- But comparing models using  $F$ -ratios isn't great if we don't really want to test null hypotheses...
- We may just want to avoid estimating more parameters than the data can support. **Fitting too many coefficients makes them all inaccurate (large SEs)**





# Information theory for model selection

In the 1970s, Hirotugu Akaike came up with a clever proof that maximum likelihood values can be “penalised” for the number of terms in a model using a value of

2

per predictor

He defined “An Information Criterion” as

$-2 * \ln(\mathcal{L}) + 2K$  where  $\mathcal{L}$  is likelihood and  $K$  is number of parameters

and showed that the model with lowest AIC is *most likely to be capable of generating the observed data*.

# Information theory for model selection

- Akaike's Information Criterion allows us to compare any number of models using a reliable measure of goodness-of-fit
- Allows for different models being almost as good → multi-model inference...
- Also *cf*  $AIC_c$  for small samples; BIC, Mallows'  $C_p$

In R: **AIC()**, **extractAIC()**

We can also use AIC for model simplification with **step()** – much easier!

# Critique of Stepwise model selection

*Journal of Animal  
Ecology* 2006  
75, 1182–1189

## Why do we still use stepwise modelling in ecology and behaviour?

MARK J. WHITTINGHAM, PHILIP A. STEPHENS\*, RICHARD B. BRADBURY† and ROBERT P. FRECKLETON‡

*Division of Biology, School of Biology and Psychology, Ridley Building, University of Newcastle, Newcastle Upon Tyne, NE1 7RU, UK; \*Department of Mathematics, University of Bristol, University Walk, Bristol, BS8 1TW, UK; †Royal Society for the Protection of Birds, The Lodge, Sandy, Bedfordshire, SG19 2DL, UK; and ‡Department of Animal and Plant Sciences, University of Sheffield, Sheffield S10 2TN, UK*

### Summary

1. The biases and shortcomings of stepwise multiple regression are well established within the statistical literature. However, an examination of papers published in 2004 by three leading ecological and behavioural journals suggested that the use of this technique remains widespread: of 65 papers in which a multiple regression approach was used, 57% of studies used a stepwise procedure.
2. The principal drawbacks of stepwise multiple regression include bias in parameter estimation, inconsistencies among model selection algorithms, an inherent (but often overlooked) problem of multiple hypothesis testing, and an inappropriate focus or reliance on a single best model. We discuss each of these issues with examples.
3. We use a worked example of data on yellowhammer distribution collected over 4 years to highlight the pitfalls of stepwise regression. We show that stepwise regression allows models containing significant predictors to be obtained from each year's data. In spite of the significance of the selected models, they vary substantially between years and suggest patterns that are at odds with those determined by analysing the full, 4-year data set.
4. An information theoretic (IT) analysis of the yellowhammer data set illustrates why the varying outcomes of stepwise analyses arise. In particular, the IT approach identifies large numbers of competing models that could describe the data equally well, showing that no one model should be relied upon for inference.

*Journal of Animal  
Ecology* 2006  
75, 1182–1189

# Example from Wittingham *et al.* (2006)

**Table 2.** Minimum adequate models constructed to explain the distribution of yellowhammers in four separate years. Data were collected from a variable number of farms in each year and these are indicated in brackets after each year

	1994 (5)	1995 (5)	1996 (8)	1997 (9)	1994–97	IT Selection probability†
Hedge presence	*	**			$P = 0.058$	0.73
Tree-line presence			*	*	***	0.67
Ditch presence	**	*		*	***	1.00
Road adjacent	*				*	0.61
Width of margin	***	*	***		***	1.00
Pasture adjacent	**		*	***	***	1.00
Silage ley adjacent						0.48
Winter rape						0.64
Beans adjacent		*				0.37
<i>n</i>	185	185	347	387	1103	
Ratio of sample size to predictors	21	21	32	35	123	

Boundary length and a code for farm forced into all models, therefore number of predictors entered into all models was 11.

\* $P < 0.05$ , \*\* $P < 0.01$ , \*\*\* $P < 0.001$ .

For comparison with the results of the full model we calculated selection probabilities using IT methodology (see Whittingham *et al.* 2005).

†The model selection probability is the probability that a given predictor will appear in the AIC-best model, and is derived from the IT-AIC analysis.

Paolo Casula

## Evaluating hypotheses about dispersal in a vulnerable butterfly

$$w_i = e^{(-\Delta_i/2)} / \sum_j e^{(-\Delta_j/2)}$$

**Table 4** Modeling survival (S), capture probability ( $p$ ) and movement probability ( $\psi$ ) in four adjacent local population of the Sardinian chalk hill blue butterfly; data coming from Supramonte di Orgosolo, Sardinia, Italy, 1999

Code	Model	Rank	AIC <sub>c</sub>	$\Delta$ AIC <sub>c</sub>	$w_i$	NP	Deviance
5bc	S(s) $p(s^*l)$ $\psi(l)$	1	1476.843	0.00	0.5809	22	369.317
6bca	S(.) $p(s^*l)$ $\psi(l)$	2	1477.881	1.04	0.3457	21	372.488
6bcb	S(s) $p(s)$ $\psi(l)$	3	1481.437	4.59	0.0584	16	386.618
4b	S( $s^*l$ ) $p(s^*l)$ $\psi(l)$	4	1484.610	7.77	0.0119	28	364.146
5ba	S( $s^*l$ ) $p(s)$ $\psi(l)$	5	1487.350	10.51	0.0030	22	379.824
4c	S(s) $p(s^*l)$ $\psi(s^*l)$	6	1497.153	20.31	0.0000	34	363.517
3	S( $s^*l$ ) $p(s^*l)$ $\psi(s^*l)$	7	1504.235	27.39	0.0000	40	357.183
6bcd	S(s) $p(s^*l)$ $\psi(.)$	8	1517.944	41.10	0.0000	11	433.542
4a	S( $s^*l$ ) $p(s^*l)$ $\psi(s)$	9	1519.930	43.09	0.0000	18	420.900
2	S( $s^*l^*t$ ) $p(s^*l)$ $\psi(s^*l)$	10	1520.805	43.96	0.0000	64	317.521
6bcc	S(s) $p(l)$ $\psi(l)$	11	1523.647	46.80	0.0000	18	424.616
5bb	S( $s^*l$ ) $p(l)$ $\psi(l)$	12	1531.169	54.33	0.0000	24	419.356
1	S( $s^*l^*t$ ) $p(s^*l^*t)$ $\psi(s^*l)$	13	1551.285	74.44	0.0000	87	289.920

Deviance is defined as the difference between  $-2\log(\text{likelihood})$  of the current model and  $-2\log(\text{likelihood})$  of the saturated model (Cooch and White 1998). NP refers to the number of estimated parameters in the current model; s, l, and t refers, respectively to sex, location, and time effects

*The likelihood ( $L$ ) of a set of parameters in a model ( $q$ ), given the data you have ( $x$ ), measures the **probability** of sampling those data given those parameters.*

# Exploring multi-variable data sets

- With lots of different variables, don't use stepwise model simplification!
- Instead, use AIC to compare models simultaneously.
- Could compare all possible subsets (e.g. R package “hier.part”) – but this is *data-dredging* – the best model out of 1024 may be no good with new data!
- Better to specify a set of candidate models and compare these.
- R package “AICcmodavg” helps you do this, and takes you through model averaging too.

# Summary

- Generate hypotheses; Design your experiment
- Plot data, Fit a model, Examine diagnostics
- Examine your model
- Designed experiments:
  - Test what you intended to test. If you didn't have hypotheses about interactions, remove them if n.s.
- Exploratory data: few variables
  - Start with maximal model
  - Simplify stepwise or use AIC
- Exploratory data: many variables
  - Specify a candidate set of models and compare using AIC

# Practical

- Linear Models Part I and Part II

## Further reading

- Burnham & Anderson (2002): Model Selection and Multimodel Inference
- Crawley: The R Book
- Whittingham *et al*: J Appl Ecol 2006