Lecture 10: Selecting Models

Bob O'Hara

bob.ohara@ntnu.no

Before we start. . .

Number of Exercises: `ceiling(N/2)`

This week's exercises don't have to be handed in until after Easter (I will out them up tomorrow)

Models

So far we have been fitting models, looking at how well they fit, and (hopefully) working out what they mean. But we also often want to compare models

- pick from a large number of models (QTLs)
- ask if one particular hypothesis explains the data

Why select models?

We could fit a model with every covariate in it

- plus all interactions

But these get big & difficult to interpret. So we only want to important variables

Smaller models generally have smaller standard errors

- more precise

Two types of problem: two solutions

Testing a specific hypothesis

- does the temperature at which a cake is baked affect how much you can break it?
- confirmatory

Finding a good model

- of these 10^6 genetic markers, which ones explain a trait?
- exploratory

Hypothesis Testing

e.g. does the temperature at which a cake is baked affect how much you can break it?

The basic approach:

- fit a model with temperature
- fit a model without temperature

Compare them & see which fits better

Hypothesis Testing

Hypothesis Testing is asymmetrical. We ask “does is model without the effect sufficient to explain the data?”

How to Do Statistical Hypothesis Testing

- get a *null hypothesis* (i.e. without the effect)
- get an *alternative hypothesis* (i.e. with the effect)
- Chose a *test statistic* (e.g. the likelihood)
- calculate the distribution of the test statistic if the null hypothesis was true
- ask if the observed value of the statistic falls within the null distribution
- if it does not, declare the null hypothesis wrong

Cake Testing

Null hypothesis: no factor explains the cake angle

Alternative hypothesis: angle can be explained by temperature

Test statistic: the likelihood

The Test statistic

We could use anything, but the likelihood makes theoretical sense

- the probability of the data given the model and parameter estimates

Need to fit the model to get the parameter estimates:

```
mod.null <- lm(angle~1, data=cake)
round(coef(mod.null), 2)
```

```
(Intercept)
      32.12
```

```
mod.alt <- lm(angle~temperature, data=cake)
round(coef(mod.alt), 2)
```

```
  (Intercept) temperature.L temperature.Q temperature.C tem
        32.12          6.61         -0.39         -0.55
temperature^5
```

The Test statistic

Question is whether the alternative model explains the data better than the null model, i.e. if the likelihood is higher

- us the ratio of likelihoods (or difference in log-likelihoods)

The likelihood is $P(Y|\hat{\theta})$ - the probability of the data given the maximum likelihood estimates of the parameters.

With maximum likelihood, the data are random. We ask

- if we had an inifinite number of samples for the data, how often would the alternative model be better than the null model
- (given the MLEs)

The Distribution of the Null Likelihood

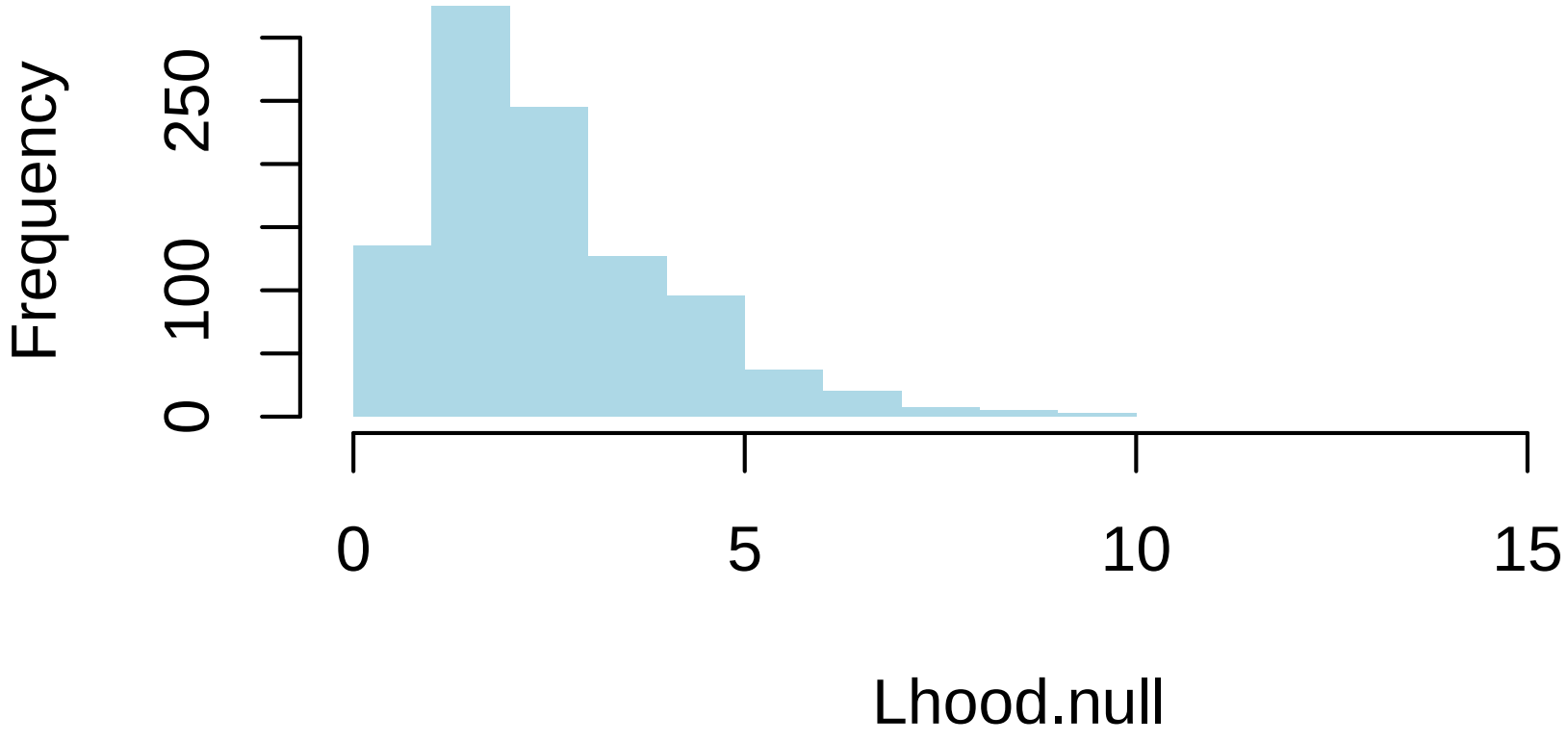
simulate() simulates data from the model in mod.null

```
CalcLhood <- function(y, x) {
  mod.null <- lm(y ~ 1)
  mod.alt <- lm(y ~ x)
  logLik(mod.alt) - logLik(mod.null)
}

# simulate the data from the (null) model
SimNull <- simulate(mod.null, nsim=1e3)
Lhood.null <- apply(SimNull, 2, CalcLhood,
                    x=cake$temperature)
```

Compare to Likelihoods

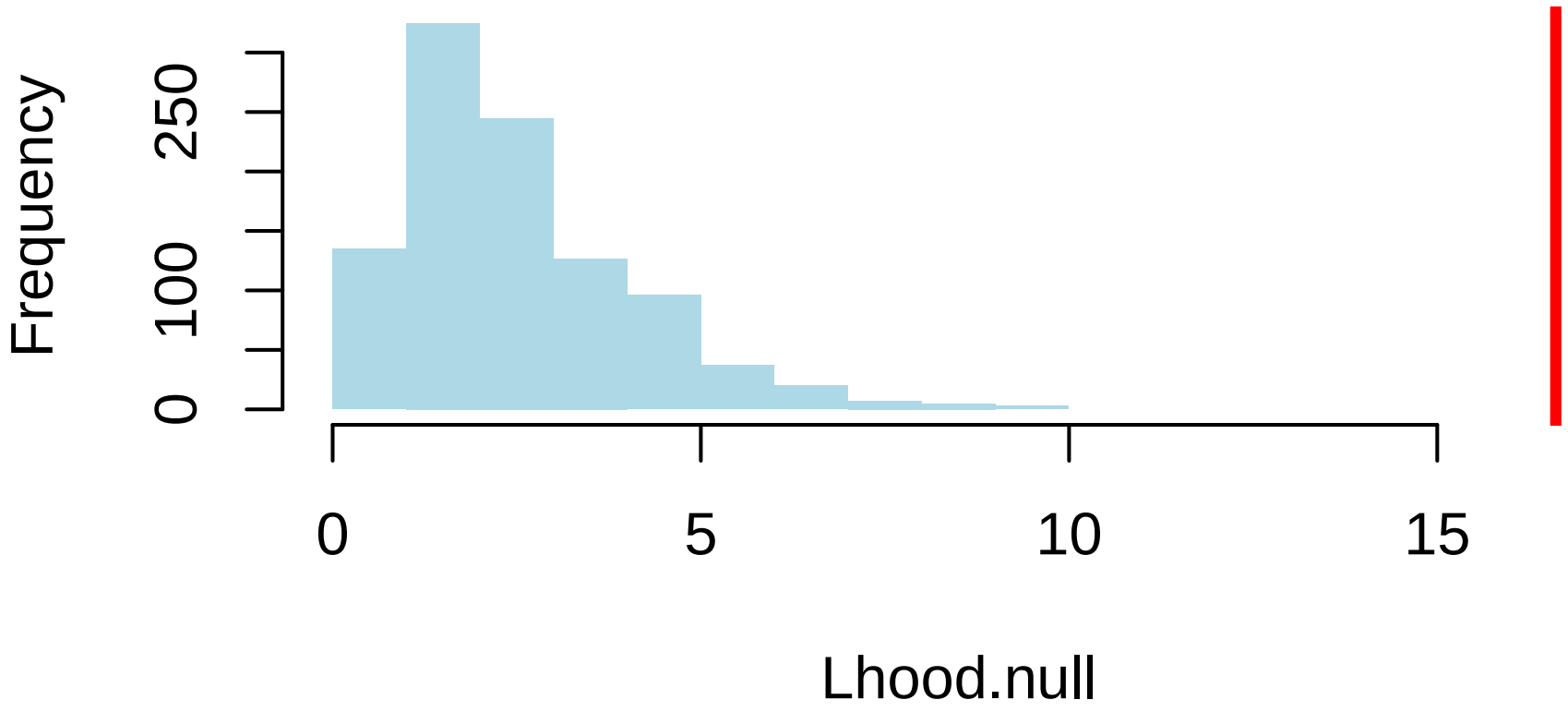
```
hist(Lhood.null, col="light blue", border=NA,
     main="", xlim=c(0, 17))
```



If temperature has an effect, the likelihood should be higher

Compare to Likelihoods

```
hist(Lhood.null, col="light blue", border=NA, main="", xlim
abline(v=CalcLhood(cake$angle, cake$temperature), col=2, lw
```



Here it is larger than all of the likelihoods

Without simulation

We know from statistical theory that the distribution of the ratio of log-likelihoods should follow an F distribution, so we don't need to simulate it

The F distribution has 2 parameters, known a "degrees of freedom".

- numerator degrees of freedom: how many extra parameters are in the alternative model
- denominator degrees of freedom: how many parameters are used to estimate $\hat{\sigma}^2$
  - taken from the alternative model

With R

We can get R to make the comparison:

```
Analysis of Variance Table

Model 1: angle ~ 1
Model 2: angle ~ temperature
  Res.Df   RSS Df Sum of Sq      F    Pr(>F)
1    269 18143
2    264 16043  5    2100.3 6.9126 4.391e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
```

The test statistic is the F-ratio (and the p-value)

Degrees of Freedom

We have N data points. Each is a "degree of freedom" that we can use in the estimation Each df can be spent to estimate one parameter The rest are used to estimate the residual variance

e.g.

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

2 parameters (α and β), so $N - 2$ can be used to estimate σ^2

$N - 2$ is the residual degrees of freedom

Degrees of Freedom

If we compare 2 models, the difference in the residual degrees of freedom is the number of extra parameters in the alternative model

- this is the degrees of freedom.

(the same as used in a χ^2 test)

With R

We can get R to make the comparison:

```
Analysis of Variance Table

Model 1: angle ~ 1
Model 2: angle ~ temperature
  Res.Df   RSS Df Sum of Sq      F    Pr(>F)
1    269 18143
2    264 16043  5    2100.3 6.9126 4.391e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
```

- df is the degrees of freedom for the model

RSS

NOT the Royal Statistical Society

The Residual Sum of Squares

- the log-likelihood for a normal distributiuon is

$$l(\mathbf{x}|\mu, \sigma^2) = -\frac{n}{2}\log\sigma^2 - \sum_{i=1}^{n}\frac{(x_i - \mu)^2}{2\sigma^2}$$

And the main bit is $\sum_{i=1}^{n}\frac{(x_i-\mu_i)^2}{2\sigma^2}$

which is just a sum of squares, and a variance term

Next

```
Analysis of Variance Table

Model 1: angle ~ 1
Model 2: angle ~ temperature
  Res.Df   RSS Df Sum of Sq      F    Pr(>F)
1    269 18143
2    264 16043  5    2100.3 6.9126 4.391e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
```

So the RSS are (almost) likelihoods

- need to divide by an estimate of σ^2
- this is the RSS for the largest model divided by its residual degrees of freedom

Why an F distribution?

-2 times difference in likelihood between 2 (nested) models follows a χ^2 distribution

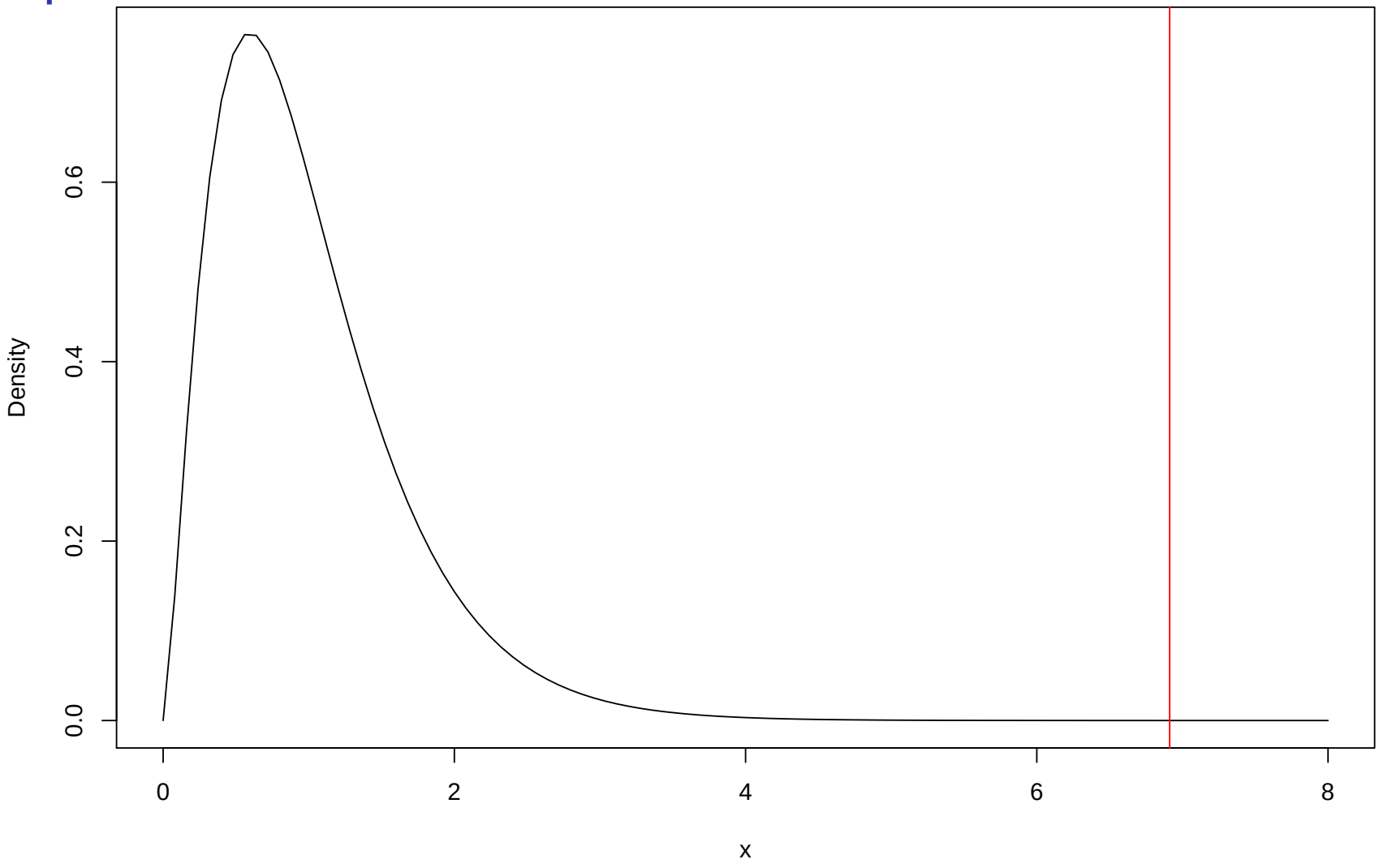
when we divide by the residual stanadrd deviation we add more uncertainty

- but this also follows a χ^2 distribution

The ratio of 2 independent χ^2 distributions (divided by their degrees of freedom) is an F distibution

- later we will see χ^2 distributions in similar tests

The p-value



The p-value is $Pr(F > F_{obs})$. Here it is NA

- so very unlikely

ANOVA made easier

We have just used anova() to compare 2 models, but it has traditionally been used to compare several:

```
Analysis of Variance Table

Response: angle
             Df  Sum Sq Mean Sq F value Pr(>F)
temperature   5  2100.3  420.06   18.27 <2e-16 ***
recipe        2   135.1   67.54    2.94   0.05 *
replicate    14 10204.2  728.87   31.69 <2e-16 ***
Residuals   248  5703.3   23.00
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
```

ANOVA made easier

Each row is a test

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
temperature	5	2100.30	420.06	18.27	0.0000
recipe	2	135.09	67.54	2.94	0.0549
replicate	14	10204.24	728.87	31.69	0.0000
Residuals	248	5703.33	23.00		

It compares a model with the terms above to one including that term e.g. the replicate line compares

```
temperature + recipe
```

to

```
temperature + recipe + replicate
```

Why is ANOVA called ANOVA

ANOVA = Analysis of Variance

The Mean Sq is the mean square, i.e. Sum Sq/Df

- when the data cooperate, it is an estimate of the variance explained by that effect
- so we can sometimes use the Mean Square to eyeball how important a variable is

Exploring for Good Models

Sometiumes we don't have strong hypotheses. Instead we might be exploring which variables might have an effect

- our aim is to get a good model overall
- e.g. for prediction

What does a good model look like?

- Simple
- Fits the data well
- Understandable

We can measure simplicity and fit.

- Fit: likelihood
- Simplicity: number of parameters

The problem

A model with always fit better if you add a parameter

Key issue: is adding the extra parameter worth it?

ANOVA answers this by asking if the improvement from the extra parameter can be explained as noise

Penalisation

Another way of looking at the problem: we measure model adqequacy

- is the model good enough for what we want?

We penalise complicated models

- measure complexity by number of parameters

Find the ‘best’ model as one with optimum between fit & complexity

How to Penalise

There are several ways to penalise. Here I will mention two, which chose different criteria

- *AIC*: Akaike's Information Criterion
- *BIC*: Bayesian Information Criterion

AIC tries to find the model that best predicts thedata

BIC trues to find the model most likely to be true

AIC

Finds the model that would best predict replicate data

AIC = -2*Likelihood + 2*Number of Parameters

BIC

Finds the model which is most likely to be “true”

BIC = -2*Likelihood + log(N)*Number of Parameters

- log(n) = log(sample size)
- penalises more than AIC

Using AIC/BIC

Full Subset Selection

- calculate AIC/BIC for every model
- pick the best

Usually, if the values are within $\sim$2 of each other, the models are pretty similar.

Example: Trying to Explain Bird Brain Size

(we will get to the actual purpose of this study in a moment)

```
BirdBrains <- read.csv("../Data/BirdBrains.csv")
BirdBrains$Mode.of.development <- factor(BirdBrains$Mode.of
BirdBrains$M.of.Dev <- c("precocial", "semi-precocial", "se
BirdBrains$M.of.Dev. <- factor(BirdBrains$M.of.Dev, levels
                                    c("altricial", "semi-altri
                                      "semi-precocial", "preco


Covars <- c("Maximum.lifespan", "Age.at.first.reprodction",
            "Incubation.length", "Clutch.size", "Mean.latit
            "logBodyMass", "M.of.Dev.")
# Scale the continuous covariates (messily)
ToScale <- Covars[sapply(Covars, function(wh, df)
  is.numeric(df[,wh]), df=BirdBrains)]
BirdBrains[,ToScale] <- scale(BirdBrains[,ToScale])
```

Look at:

Fit the model

```
library(bestglm) # might need install.packages("bestglm")
UseData <- cbind(BirdBrains[,Covars], y=BirdBrains$logBrain
AllSubsets <- bestglm(Xy=UseData, IC="AIC")
```

```
Morgan-Tatar search since factors present with more than 2
```

```
AllSubsets
```

```
AIC
Best Model:
                  Df Sum Sq Mean Sq F value Pr(>F)
Maximum.lifespan   1 249.72  249.72 3852.11 <2e-16 ***
logBodyMass        1 180.81  180.81 2789.06 <2e-16 ***
M.of.Dev.          3  12.08    4.03   62.09 <2e-16 ***
Residuals        378  24.50    0.06
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
```

The Best model

Max. lifespan, body mass, mode of development (how developed a chick is when it hatches: precocial is most developed)

Model has R^2 of 95

	Estimate	Std. Error	t value	Pr(>|t
(Intercept)	0.70	0.02	37.29	0.(
Maximum.lifespan	0.08	0.02	3.98	0.(
logBodyMass	1.10	0.02	48.23	0.(
M.of.Dev.semi-altricial	-0.02	0.06	-0.37	0.7
M.of.Dev.semi-precocial	-0.32	0.05	-5.81	0.(
M.of.Dev.precocial	-0.48	0.04	-12.20	0.(

Why we should't just use anova() Pt 1

```
anova(lm(logBrainMass~Mean.latitude+logBodyMass+Maximum.li
         data=BirdBrains))
```

```
Analysis of Variance Table

Response: logBrainMass
                  Df Sum Sq Mean Sq   F value    Pr(>F)
Mean.latitude      1   0.63    0.63    9.7624  0.001919 **
logBodyMass        1 429.77  429.77 6629.7470 < 2.2e-16 ***
Maximum.lifespan   1   1.67    1.67   25.7511 6.111e-07 ***
M.of.Dev           3  10.60    3.53   54.5278 < 2.2e-16 ***
Residuals        377  24.44    0.06
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
```

Mean latitude has an effect!

Why we should't just use anova() Pt 2

```
anova(lm(logBrainMass~Maximum.lifespan+M.of.Dev+logBodyMass
         data=BirdBrains))
```

```
Analysis of Variance Table

Response: logBrainMass
                   Df  Sum Sq Mean Sq  F value Pr(>F)
Maximum.lifespan    1 249.724 249.724 3852.321 <2e-16 ***
M.of.Dev            3  42.085  14.028  216.406 <2e-16 ***
logBodyMass         1 150.799 150.799 2326.284 <2e-16 ***
Mean.latitude       1   0.066   0.066    1.021 0.3129
Residuals         377  24.439   0.065
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
```

Mean latitude has no effect!

Why we should't just use anova()

Because the tests are sequential, the order usually matters

(unless the data are from a well-designed experiment)

When to Use these Methods

ANOVA: testing specific hypotheses

AIC/Full subsets: finding the best model

Bird Brains: how I would approach it

For the bird brains, the data were collected to ask if life span was related to brain size. Other variables were included because they might have an effect (e.g. body size)

So, we have a spcific hypothesis to test

But we want to find the other covaraites that might have an effect

So, we use AIC to find these other covariates

Bird Brains: how I would approach it

```
# Don't use lifespan here, but select best model with the
UseData <- cbind(BirdBrains[,Covars[-1]], y=BirdBrains$logB
AllSubsets <- bestglm(Xy=UseData, IC="AIC")
```

Morgan-Tatar search since factors present with more than 2

```
# Now add max. lifespan
TestLife <- update(AllSubsets$BestModel,
                   .~.+Maximum.lifespan, data=BirdBrains)
```

The Result

```
anova(AllSubsets$BestModel, TestLife) # could use anova(Tes
```

```
Analysis of Variance Table

Model 1: y ~ Age.at.first.reprodction + logBodyMass + M.of
Model 2: y ~ Age.at.first.reprodction + logBodyMass + M.of
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    378 25.140
2    377 24.425  1   0.71441 11.027 0.0009857 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
```

Yep, it still has an effect

The Model

. . . but the effect is small compared to the other effects (note: these are standardised)

```
round(summary(TestLife)$coefficients,2)
```

```
                                Estimate Std. Error t value Pr(>|t
(Intercept)                         0.70       0.02   37.21      0
Age.at.first.reprodction            0.02       0.02    1.11      0
logBodyMass                         1.09       0.02   45.40      0
M.of.Dev.semi-altricial            -0.03       0.06   -0.47      0
M.of.Dev.semi-precocial            -0.35       0.06   -5.66      0
M.of.Dev.precocial                 -0.48       0.04  -12.04      0
Maximum.lifespan                    0.07       0.02    3.32      0
```

Next Week

I'll be jealous of you in Borneo

After Easter, I will wrap up where we are & how everything fits together

The exercises don't have to be handed in until after Easter.