

Statistical inference and linear models

February 22, 2018

If you get bored

- Go to the last slide for bonus exercises
- Work on code for your research and ask question during exercise time
- But try and keep an eye out, just in case



Disclaimer

- Assume you got lectures about statistics and
 - ▶ know why we need statistics
 - ▶ have heard of the general philosophy
- We may simplify to focus on practical aspects
- Correct us if we say something completely awful

- 1 Statistical inference
- 2 t-test, ANOVA, regression: all is one, one is all
- 3 Linear models in details
- 4 Bonus fun

General approach

1. Scientific question

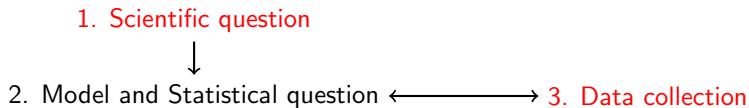
General approach

1. Scientific question

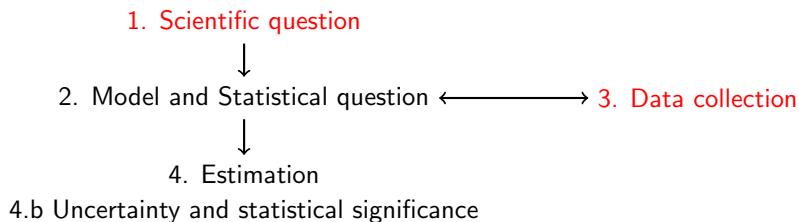


2. Model and Statistical question

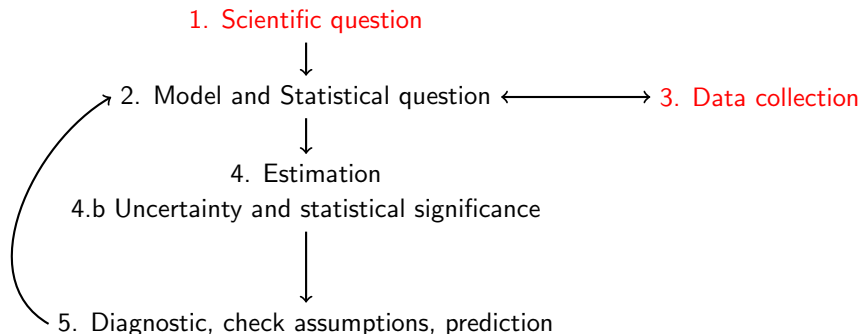
General approach



General approach

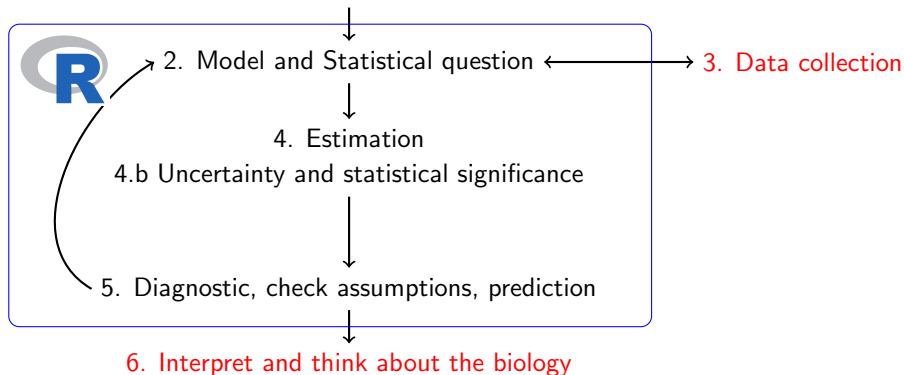


General approach



General approach

1. Scientific question



Reminder t.test

Reminder t.test

```
data("iris")
```

```
str(iris)  
plot(iris)
```

- 1 Scientific question: Are the taxa "setosa" and "versicolor" different species?

Reminder t.test

```
data("iris")
```

```
str(iris)  
plot(iris)
```

- ① Scientific question: Are the taxa "setosa" and "versicolor" different species?
- ② Model and stat question:
 - ▶ Model:
 - ★ There is an intrinsic/expected sepal length value for a species; an individual value is the sum of this expectation and a random Gaussian deviation.
 - ★ $y_i = \mu_{species_i} + \epsilon_i$ with $\epsilon \sim N(0, \sigma^2)$
 - ★ t-test

Reminder t.test

```
data("iris")
```

```
str(iris)  
plot(iris)
```

① Scientific question: Are the taxa "setosa" and "versicolor" different species?

② Model and stat question:

► Model:

- ★ There is an intrinsic/expected sepal length value for a species; an individual value is the sum of this expectation and a random Gaussian deviation.
- ★ $y_i = \mu_{species_i} + \epsilon_i$ with $\epsilon \sim N(0, \sigma^2)$
- ★ t-test

► Statistical question:

- ★ Does sepal length **differ significantly** between the two taxa **in our sample**?
- ★ Is the observed difference between taxa likely if both taxa have the same intrinsic/expected value?

③ Data collection

Reminder t.test

One t-test for sepal length between *setosa* and *versicolor*:

```
t.test(x = iris$Sepal.Length[iris$Species == "setosa"],  
       y = iris$Sepal.Length[iris$Species == "versicolor"])
```

Welch Two Sample t-test

```
data:  iris$Sepal.Length[iris$Species == "setosa"] and iris$Sepal.Length[iris$Species == "versicolor"]  
t = -10.521, df = 86.538, p-value < 2.2e-16  
alternative hypothesis: true difference in means is not equal to 0  
95 percent confidence interval:  
 -1.1057074 -0.7542926  
sample estimates:  
mean of x mean of y  
   5.006    5.936
```

When do we know it is different?

④ Statistical estimation

► a Estimation

- ★ Cannot know true difference $\mu_{species_1} - \mu_{species_2}$
- ★ Estimated difference = **Mean₁ - Mean₂**
- ★ Difference contains random variation

► b Quantify uncertainty / Statistical significance

- ★ $t = \frac{\text{Mean}_1 - \text{Mean}_2}{\text{Variation}} \frac{\sqrt{\text{Sample Size}}}{\sqrt{2}}$
- ★ We know exactly how t is distributed when $\mu_{species_1} - \mu_{species_2} = 0$
- ★ Hence we know probability of $\geq t$ if $\mu_{species_1} - \mu_{species_2} = 0$ (p -value)
- ★ Can derive confidence interval and standard error

When do we know it is different?

• Statistical estimation

▶ a Estimation

- ★ Cannot know true difference $\mu_{species_1} - \mu_{species_2}$
- ★ Estimated difference = **Mean₁ - Mean₂**
- ★ Difference contains random variation

▶ b Quantify uncertainty / Statistical significance

- ★ $t = \frac{\text{Mean}_1 - \text{Mean}_2}{\text{Variation}} \frac{\sqrt{\text{Sample Size}}}{\sqrt{2}}$
- ★ We know exactly how t is distributed when $\mu_{species_1} - \mu_{species_2} = 0$
- ★ Hence we know probability of $\geq t$ if $\mu_{species_1} - \mu_{species_2} = 0$ (p -value)
- ★ Can derive confidence interval and standard error

Less uncertainty with

- **Larger absolute difference**
- **Smaller variability**
- **Larger sample size**

When do we know it is different? Simulations

When do we know it is different? Simulations

1. Larger absolute difference

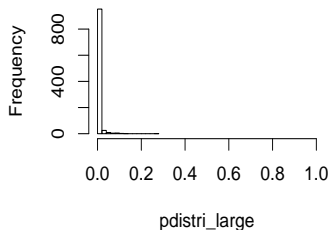
```
nbsim <- 1000
pdistri_large <- vector(length = nbsim)
pdistri_small <- vector(length = nbsim)
for (i in 1:nbsim)
{
  x1 <- rnorm(n = 10, mean = 2, sd = 1)
  x2 <- rnorm(n = 10, mean = 4, sd = 1) #large diff
  x3 <- rnorm(n = 10, mean = 2.5, sd = 1) #small diff
  out_large <- t.test(x1, x2)
  out_small <- t.test(x1, x3)
  pdistri_large[i] <- out_large$p.value
  pdistri_small[i] <- out_small$p.value
}
```

How would you visualize these two vectors?

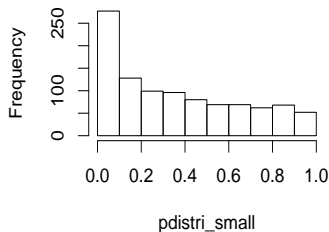
When do we know it is different?

```
par(mfrow=c(1,2), cex=2)
hist(pdistrib_large, xlim=c(0,1),
     main=paste("Prop signif=",mean(pdistrib_large<0.05)))
hist(pdistrib_small, xlim=c(0,1),
     main=paste("Prop signif=",mean(pdistrib_small<0.05)))
```

Prop signif= 0.983



Prop signif= 0.179



```
par(mfrow=c(1,1))
```

When do we know it is different? Try it!

Exercise

Check the effect of **smaller variability** and/or **larger sample size**.

By the way, what are these p-values?

*Probability for a summary statistic to be greater or equal to the observed summary statistic, **when the null-hypothesis of a given statistical model is true.***

By the way, what are these p-values?

*Probability for a summary statistic to be greater or equal to the observed summary statistic, **when the null-hypothesis of a given statistical model is true.***

Properties

- Depends on the null-hypothesis (H_0) of a given model with assumptions
- Uniform distribution under H_0 ...
- ... hence $\text{proportion}(\text{significance under } H_0) = \text{significance threshold}$

By the way, what are these p-values?

*Probability for a summary statistic to be greater or equal to the observed summary statistic, **when the null-hypothesis of a given statistical model is true.***

Properties

- Depends on the null-hypothesis (H_0) of a given model with assumptions
- Uniform distribution under H_0 ...
- ...hence $\text{proportion}(\text{significance under } H_0) = \text{significance threshold}$

```
hist( runif(n = 10000, min = 0, max=1) )
```


By the way, what are these p-values?

*Probability for a summary statistic to be greater or equal to the observed summary statistic, **when the null-hypothesis of a given statistical model is true.***

Properties

- Depends on the null-hypothesis (H_0) of a given model with assumptions
- Uniform distribution under H_0 ...
- ...hence $\text{proportion}(\text{significance under } H_0) = \text{significance threshold}$

```
hist( runif(n = 10000, min = 0, max=1) )
```

NB: Focus on p -value criticized, but common and they are no more evil than other misused statistics!

T-test exercise: p-values and simulations

```
t.test(x = ..., y=..., var.equal = TRUE)
t.test(x = ..., y=..., var.equal = FALSE)
```

What if variances are different by chance only?

```
set.seed(1234)
var(rnorm(20, mean = 0, sd = 1))
```

```
[1] 1.027806
```

```
var(rnorm(20, mean = 0, sd = 1))
```

```
[1] 0.6265501
```

Exercise

What option is more correct for var.equal?

- 1 Statistical inference
- 2 t-test, ANOVA, regression: all is one, one is all
- 3 Linear models in details
- 4 Bonus fun

A small example

Animal behavior in response to weather

Load data:

```
getwd()  
setwd()
```

```
dat.behav <- read.csv(file = "datbehav.csv") # path to file
```

A small example

Animal behavior in response to weather

Load data:

```
getwd()  
setwd()
```

```
dat.behav <- read.csv(file = "datbehav.csv") # path to file
```

STEP 1: have a look at your data

```
str(dat.behav)  
summary(dat.behav)  
plot(dat.behav)
```

t-test

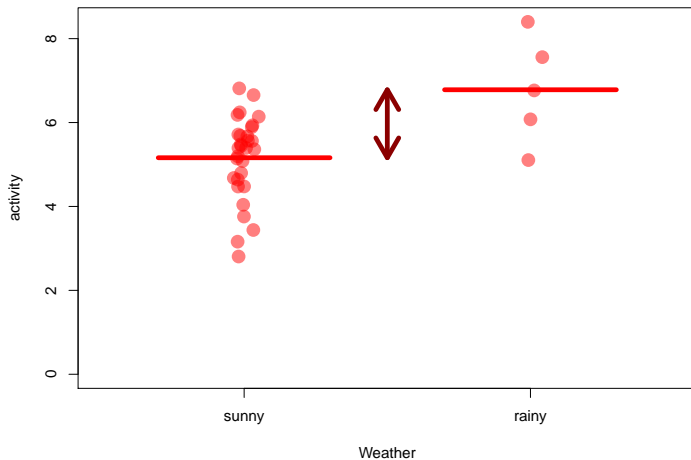
```
fitstudent <- t.test(x = dat.behav$activity[dat.behav$weather=="rainy"],
                    y = dat.behav$activity[dat.behav$weather=="sunny"],
                    var.equal = TRUE)
print(fitstudent)
```

Two Sample t-test

```
data: dat.behav$activity[dat.behav$weather == "rainy"] and dat.behav$activity[dat.behav$weather == "sunny"]
t = 3.2752, df = 33, p-value = 0.002485
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.6138373 2.6270325
sample estimates:
mean of x mean of y
 6.781476  5.161041
```

t-test, graphically

Difference between means



ANOVA

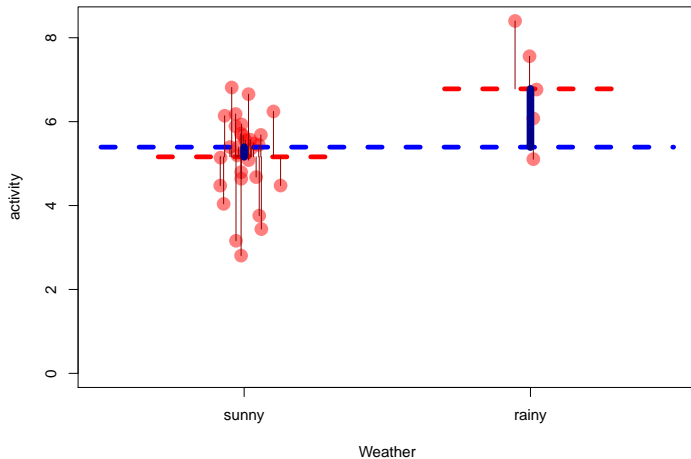
```
fitanova <- aov(data = dat.behav, formula = activity ~ weather)
summary(fitanova)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
weather	1	11.25	11.253	10.73	0.00248 **
Residuals	33	34.62	1.049		

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

ANOVA, graphically

Variance decomposition



Linear regression

```
fitlm <- lm(data = dat.behav, formula = activity ~ weather)
summary(fitlm)
```

Call:

```
lm(formula = activity ~ weather, data = dat.behav)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-2.3547	-0.6028	0.2346	0.6419	1.6534

Coefficients:

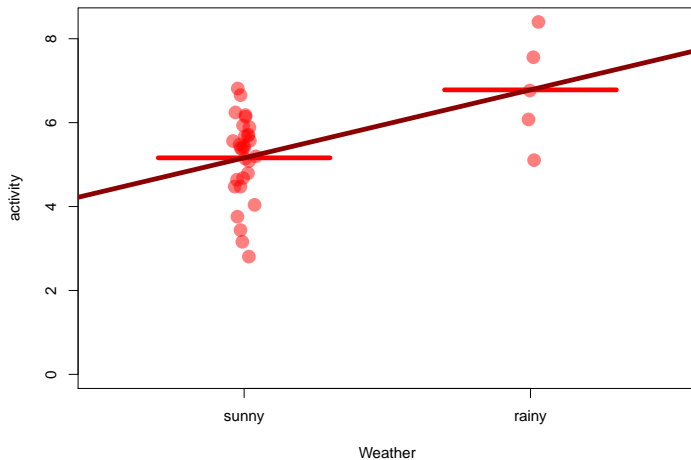
	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	6.7815	0.4581	14.805	3.94e-16 ***
weathersunny	-1.6204	0.4948	-3.275	0.00248 **

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

Residual standard error: 1.024 on 33 degrees of freedom

Regression, graphically

Rate of change



NB: aov() vs. anova()

```
aov(data = dat.behav, formula = activity ~ weather)
anova(fitlm)
```

All is one. . .

All is one. . .

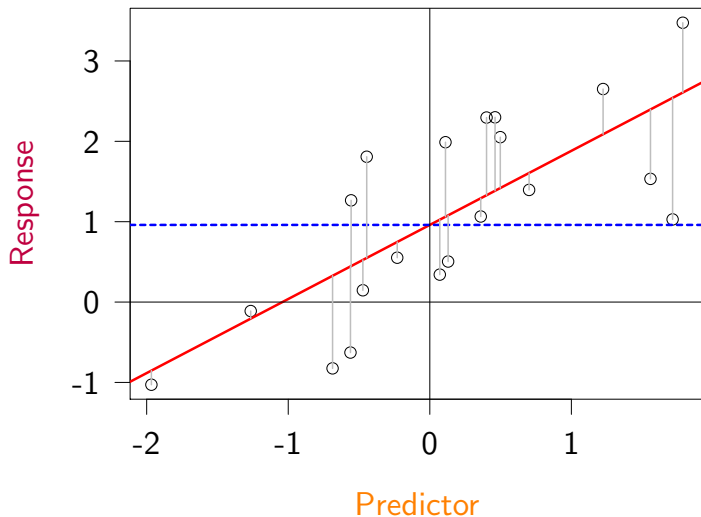
...but `lm()` rules!

- t-test, ANOVA, regression and others can be mathematically equivalent
- In R, `lm()` and related functions can do them all. . .
- ...and much more!

- 1 Statistical inference
- 2 t-test, ANOVA, regression: all is one, one is all
- 3 Linear models in details**
- 4 Bonus fun

A simple linear model

$$\text{Response} = \text{Intercept} + \text{Slope} \times \text{Predictor} + \text{Error}$$



A simple linear model

$$\text{Response} = \text{Intercept} + \text{Slope} \times \text{Predictor} + \text{Error}$$

In R:

```
lm(response ~ 1 + predictor1 + predictor2, data=data)
# equivalent to
lm(response ~ predictor1 + predictor2, data=data)
```

- Intercept can be explicit or implicit
- Can remove intercept with $\dots \sim 0 + \dots$
- Error is implicit
- Feed the option `data=` to keep code short, reliable and flexible
- Order of predictors do not matter

Interpretation

```
Ans <- read.csv(file = "Anscombe.csv")
```

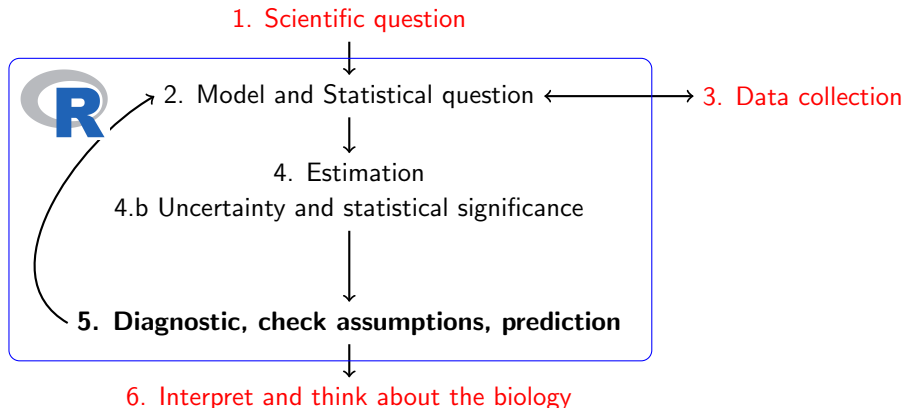
```
lm1 <- lm(y ~ x , data=Ans[Ans$distri==1,])  
summary(lm1)  
plot(Ans$x[Ans$distri==1], Ans$y[Ans$distri==1],  
      xlim=c(0,15), ylim=c(0,12))  
abline(lm1)
```

Interpretation

lm vs. plot

- Fit a linear model $y \sim x$ for each of the four “distri”
- Plot the relationship $y \sim x$ for each of the four “distri”
- Can we trust these models? For what? *I expect more than “it’s all bullshit”*

General approach



Linear model basic assumptions

Not necessarily wrong, but typical interpretation assumes:

- Linear combination of parameters (including transformation, polynoms, interactions. . .)
Risk: biologically meaningless

Linear model basic assumptions

Not necessarily wrong, but typical interpretation assumes:

- Linear combination of parameters (including transformation, polynoms, interactions. . .)
Risk: biologically meaningless
- Predictor not perfectly correlated
Risk: Model won't run, unstable convergence, or huge SE

Linear model basic assumptions

Not necessarily wrong, but typical interpretation assumes:

- Linear combination of parameters (including transformation, polynoms, interactions. . .)
Risk: biologically meaningless
- Predictor not perfectly correlated
Risk: Model won't run, unstable convergence, or huge SE
- Little error in predictors
Risk: bias estimates (underestimate with Gaussian error)

Linear model basic assumptions

Not necessarily wrong, but typical interpretation assumes:

- Linear combination of parameters (including transformation, polynoms, interactions. . .)
Risk: biologically meaningless
- Predictor not perfectly correlated
Risk: Model won't run, unstable convergence, or huge SE
- Little error in predictors
Risk: bias estimates (underestimate with Gaussian error)
- Gaussian error distribution
Risk: Poor predictions

Linear model basic assumptions

Not necessarily wrong, but typical interpretation assumes:

- Linear combination of parameters (including transformation, polynoms, interactions. . .)
Risk: biologically meaningless
- Predictor not perfectly correlated
Risk: Model won't run, unstable convergence, or huge SE
- Little error in predictors
Risk: bias estimates (underestimate with Gaussian error)
- Gaussian error distribution
Risk: Poor predictions
- Homoscedasticity (constant error variance)
Risk: Over-optimistic uncertainty, unreliable predictions

Linear model basic assumptions

Not necessarily wrong, but typical interpretation assumes:

- Linear combination of parameters (including transformation, polynoms, interactions. . .)
Risk: biologically meaningless
- Predictor not perfectly correlated
Risk: Model won't run, unstable convergence, or huge SE
- Little error in predictors
Risk: bias estimates (underestimate with Gaussian error)
- Gaussian error distribution
Risk: Poor predictions
- Homoscedasticity (constant error variance)
Risk: Over-optimistic uncertainty, unreliable predictions
- Independence of error
Risk: Bias and over-optimistic uncertainty

Diagnostic: summary and plot

```
lm1 <- lm(y ~ x , data=Ans[Ans$distri==1,])  
lm2 <- lm(y ~ x , data=Ans[Ans$distri==2,])
```

```
summary(lm1)  
par(mfrow=c(2,2))  
plot(lm1)
```

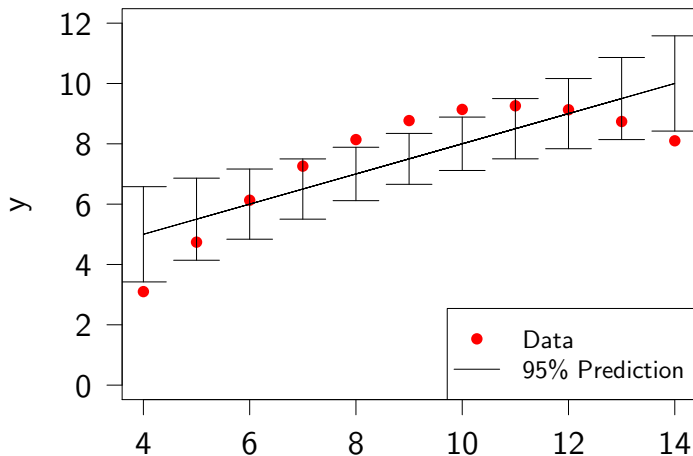
```
summary(lm2)  
plot(lm2)  
par(mfrow=c(1,1))
```

Diagnostic: prediction

```
pred2 <- predict(lm2, se.fit = TRUE, interval = "confidence")  
pred2 <- cbind(Ans[Ans$distri==2,], pred2)
```

Diagnostic: prediction

```
pred2 <- predict(lm2, se.fit = TRUE, interval = "confidence")  
pred2 <- cbind(Ans[Ans$distri==2,], pred2)
```



Practice lm() with parasites

What explains variation in parasitic load?

You collected ecto-parasites on some furry large mammals at three locations. Parasites break easily when we collect them and are impossible to count, so we decide to measure parasitic load as their mass. **Why do some mammals have larger parasitic load?**

Practice `lm()` with parasites

What explains variation in parasitic load?

You collected ecto-parasites on some furry large mammals at three locations. Parasites break easily when we collect them and are impossible to count, so we decide to measure parasitic load as their mass. **Why do some mammals have larger parasitic load?**

- Load the `Para.csv` data (don't forget: `str()`, `summary()`, `plot()`...)
- Model `Parasite_Mass` using `lm()`
- Find what variables predict `Parasite_Mass`
- How good are your models? Assumptions? Prediction?
- What biological interpretation can you imagine?

- 1 Statistical inference
- 2 t-test, ANOVA, regression: all is one, one is all
- 3 Linear models in details
- 4 Bonus fun

Extra exercises

General R coding

- 1 What is the fastest way to get row averages in a data-frame?
- 2 Create a function called `colVars`, like `colMeans` but for variance
- 3 Create nice plots to visualize iris data (ideally journal-quality)

Linear models

- 1 Load `Cdata.csv`, fit models of y predicted by x_1 and x_2 , or x_2 and x_3 . Something is weird, what is going on? What to do?
- 2 For model that can be fitted with `t.test`, `aov`, and `lm`, is one of the function faster?
- 3 Write your own code to obtain a prediction from a `lm` (that is, a simpler version of the `predict` function), with confidence interval. (extra toughness: do it using the matrix formulation of the analytical solution to a linear model)