Likelihood and all that

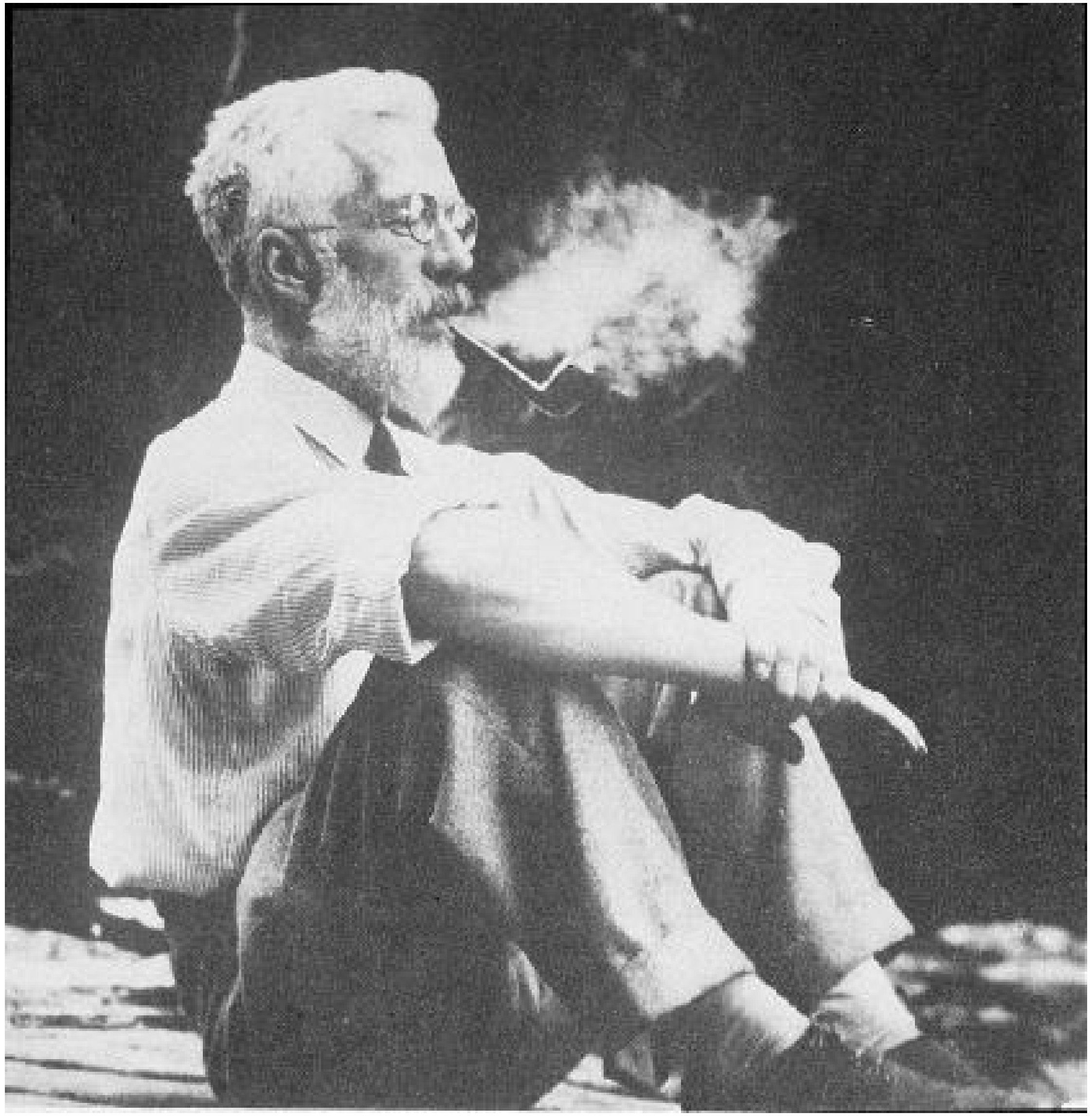
Anders Nielsen, Ethan Lawler, & Sean Anderson

an@aqua.dtu.dk

Outline

Don't worry this will not turn into a statistics course, but just a gentle reminder of

- Likelihood function $L(\theta) = P_\theta(Y = y)$
- Negative log likelihood function $\ell(\theta) = -\log(L(\theta))$
- Maximum likelihood estimate $\widehat{\theta} = \operatorname*{argmin}_{\theta \in \Theta} \ell(\theta)$
- Distribution of the ML estimator $\widehat{\theta} \sim \mathrm{N}(\theta, (\ell''(\widehat{\theta}))^{-1})$
- Likelihood ratio test $2(\ell_B(\widehat{\theta_B}, Y) - \ell_A(\widehat{\theta_A}, Y)) \sim \chi^2_{\dim(A)-\dim(B)}$



Sir Ronald Aylmer Fisher (1890 - 1962) identified the likelihood function as the key inferential quantity conveying all inferential information in statistical modelling including the uncertainty.

Observations with noise

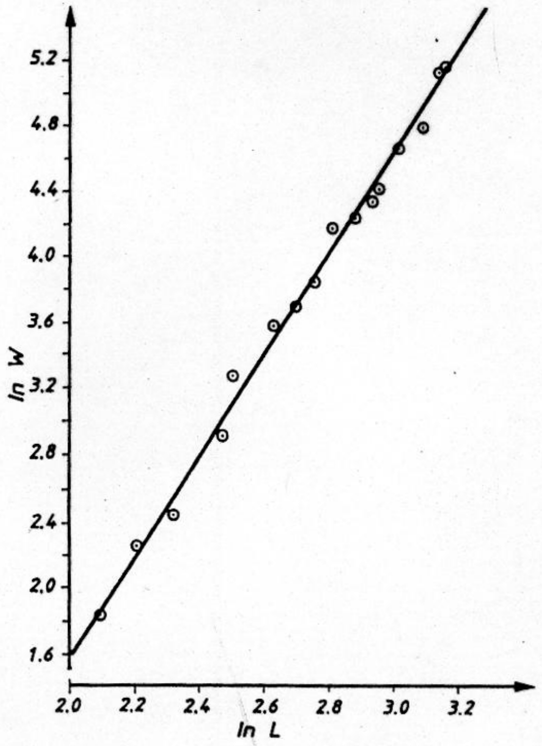


Fig. 2.6.2 The data from Fig. 2.6.1 converted to natural logarithms

- “Noise” is slang for unexplained variation in our observations
- Here the model $\log W_i = \alpha + \beta \log L_i$ is a very good description
- Still something is missing, as all the points are not exactly on the line

Statistical models

- Statistical models are explicit about the noise term.

$$\log W_i = \alpha + \beta \log L_i + \varepsilon_i, \text{ where } \varepsilon_i \sim \mathcal{N}(0, \sigma^2) \text{ independent}$$

- Because:
  - We want to explain the entire system
  - We can better explain how good our model is
  - It help us to estimate the model parameters
  - It help us to quantify uncertainties on model parameters
  - It gives us an objective criteria for comparing models
  - ...
- Example question: Let's say I have a fish with log-length of 3.0 what can we say about its log-weight?
- Example question: How certain are we about our slope estimate?

Biological data uncertain?

- Catch at age data for instance:
  - Weights of (officially) landed fish
  - Samples of lengths
  - Samples of ages
  - Estimates of amount discarded at sea
- What do you think?

Maximum likelihood estimation

- This is the general approach to estimating model parameters
- It gives a complete recipe for estimating parameters $\hat{\theta} = \text{argmin}_{\theta}\ell(\theta|\texttt{data})$
- In practice this often boils down to:

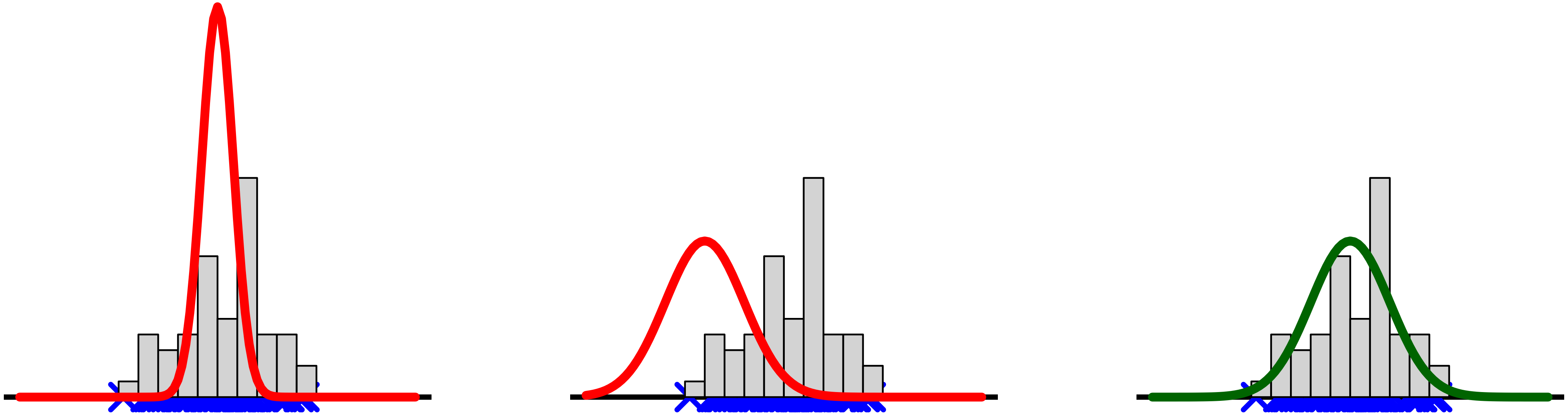
  1. Setup a function $\ell(\theta|\texttt{data})$ to calculate the negative log-likelihood of the entire data set acording to the model when the model parameters are θ
  2. Assign starting value to θ
  3. Use an iterative function minimizer (e.g. `nlminb`) to find the minimum

Principle is logical

- We have:

  Observations: $y = (y_1, y_2, \dots, y_n)$

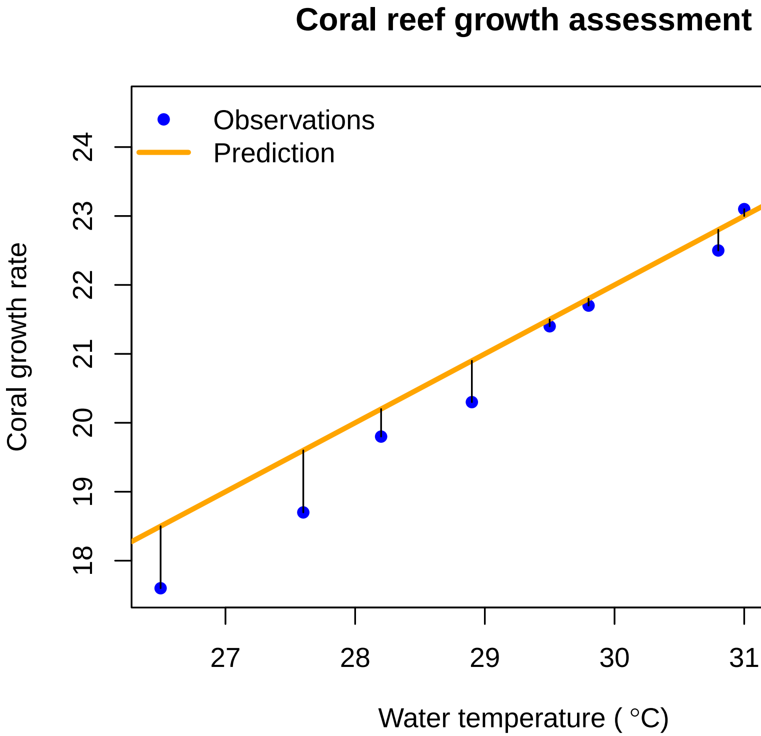
  Parameters (μ, σ) in model: $y_i \sim N(\mu, \sigma^2)$



- Choose parameters which makes our model best match the data (optimize likelihood).

Consider e.g. a linear regression example

Water temperature (x)	Coral growth rate (y)
29.5	21.4
31.0	23.1
28.2	19.8
30.8	22.5
27.6	18.7
32.1	24.6
29.8	21.7
28.9	20.3
26.5	17.6
31.3	23.9



- The model is:

$$y_i = \alpha x_i + \beta + \varepsilon_i \quad , \quad \text{where } \varepsilon_i \sim \mathbf{N}(0, \sigma^2) \text{ independently,} \quad i = 1, \ldots, n$$

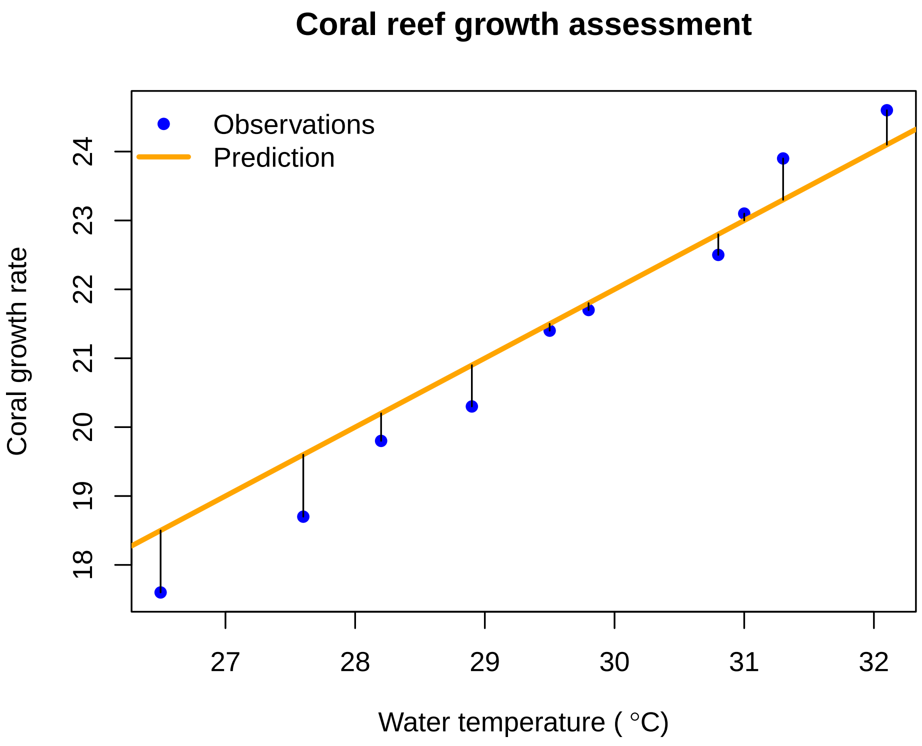
- If we assign a value to our model parameters e.g: $\theta = (\alpha = 1, \beta = -8, \text{ and } \sigma = 2)$, then we can calculate the likelihood of the first observation as:

```
> dnorm(21.4,mean=1*29.5-8,sd=2) # 0.199222
```

Mini exercise: What is the likelihood of the second observation? What is the joint likelihood of the two first observations?

… and the likelihood

Water temperature (x)	Coral growth rate (y)
29.5	21.4
31.0	23.1
28.2	19.8
30.8	22.5
27.6	18.7
32.1	24.6
29.8	21.7
28.9	20.3
26.5	17.6
31.3	23.9



- The model is:

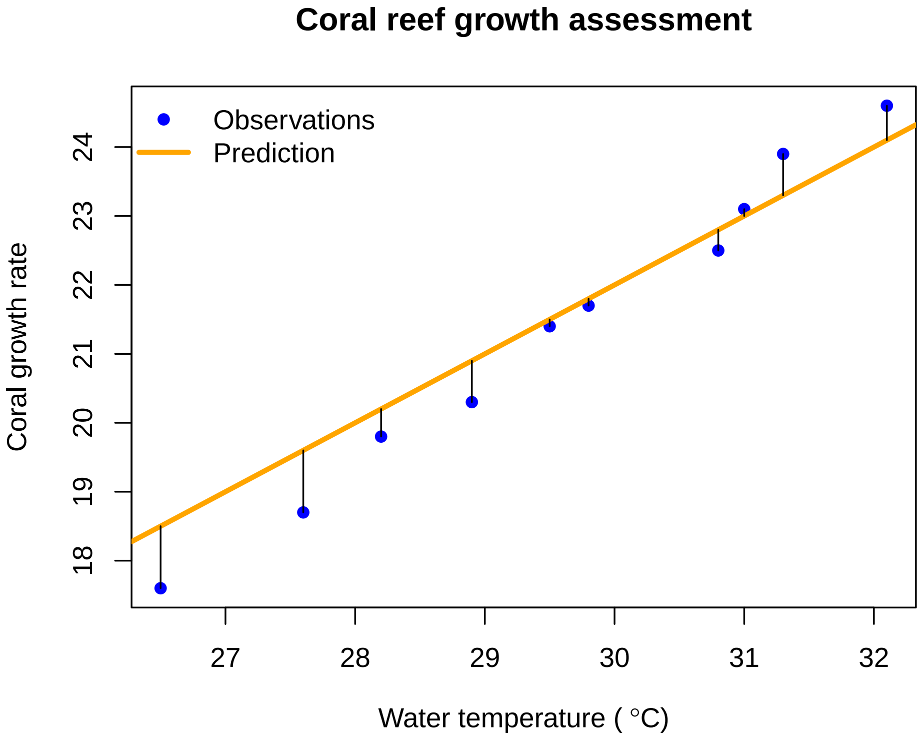
$$y_i = \alpha x_i + \beta + \varepsilon_i \quad , \quad \text{where } \varepsilon_i \sim \mathbf{N}(0, \sigma^2) \text{ independently}, \quad i = 1, \ldots, n$$

- We can calculate the likelihood of all 10 observations by:

```
> x <- c(29.5, 31.0, 28.2, 30.8, 27.6, 32.1, 29.8, 28.9, 26.5, 31.3)
> y <- c(21.4, 23.1, 19.8, 22.5, 18.7, 24.6, 21.7, 20.3, 17.6, 23.9)
> L <- function(th){ prod(dnorm(y, mean=th[1]*x+th[2], sd=th[3]))}
> L(c(1,-8,2))  #  6.966226e-08
```

… and the negative log likelihood

Water temperature (x)	Coral growth rate (y)
29.5	21.4
31.0	23.1
28.2	19.8
30.8	22.5
27.6	18.7
32.1	24.6
29.8	21.7
28.9	20.3
26.5	17.6
31.3	23.9



- The model is:

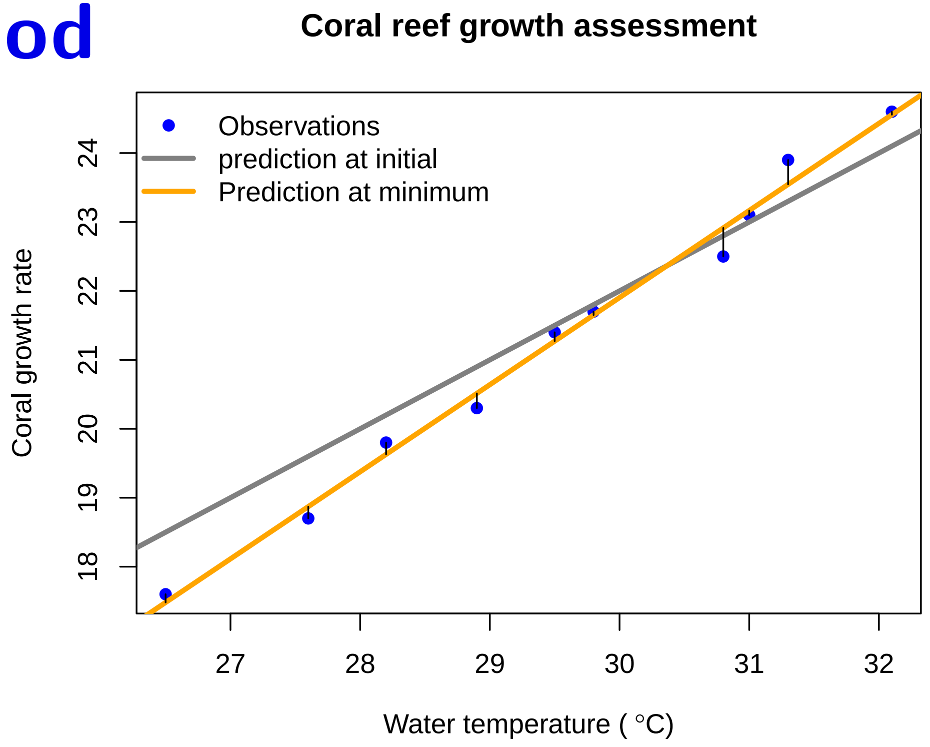
$$y_i = \alpha x_i + \beta + \varepsilon_i \quad , \quad \text{where } \varepsilon_i \sim \mathbf{N}(0, \sigma^2) \text{ independently}, \quad i = 1, \ldots, n$$

- We can calculate the likelihood of all 10 observations by:

```
> l <- function(th){ -sum(dnorm(y, mean=th[1]*x+th[2], sd=th[3], log=TRUE))}
> l(c(1,-8,2))  #  16.47961
```

… and minimize the negative log likelihood

Water temperature (x)	Coral growth rate (y)
29.5	21.4
31.0	23.1
28.2	19.8
30.8	22.5
27.6	18.7
32.1	24.6
29.8	21.7
28.9	20.3
26.5	17.6
31.3	23.9



- The model is:

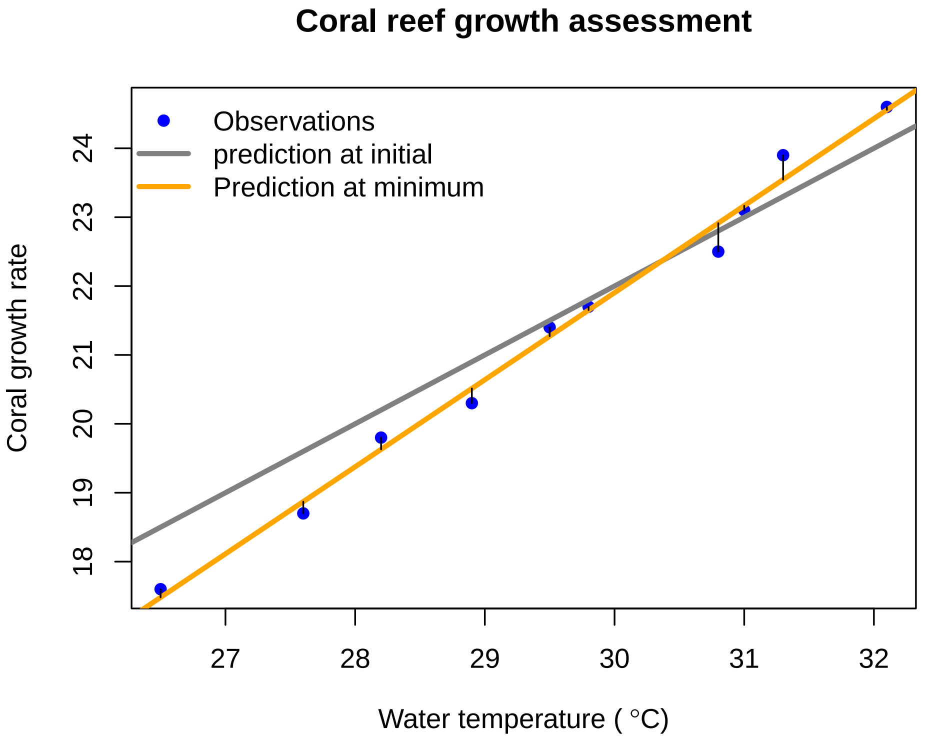
$$y_i = \alpha x_i + \beta + \varepsilon_i \quad , \quad \text{where } \varepsilon_i \sim \mathbf{N}(0, \sigma^2) \text{ independently}, \quad i = 1, \ldots, n$$

- We can calculate the likelihood of all 10 observations by:

```
> l <- function(th){ -sum(dnorm(y, mean=th[1]*x+th[2], sd=exp(th[3]), log=TRUE))}
> fit<-nlminb(c(alpha=1,beta=-8, logSigma=0), l)
> abline(fit$par[2:1], lwd=3, col="orange")
> exp(fit$par[3])                                # sigma ca. 0.21
> arrows(x,y,x,fit$par[1]*x+fit$par[2], code=0)
```

… and same via RTMB

Water temperature (x)	Coral growth rate (y)
29.5	21.4
31.0	23.1
28.2	19.8
30.8	22.5
27.6	18.7
32.1	24.6
29.8	21.7
28.9	20.3
26.5	17.6
31.3	23.9



- The model is:

$$y_i = \alpha x_i + \beta + \varepsilon_i \quad , \quad \text{where } \varepsilon_i \sim \mathbf{N}(0, \sigma^2) \text{ independently}, \quad i = 1, \ldots, n$$

- We can calculate the likelihood of all 10 observations by:

```
> library(RTMB)
> par <- list(alpha=1, beta=-8, logSigma=0)
> l <- function(par){ -sum(dnorm(y, mean=par$alpha*x+par$beta, sd=exp(par$logSigma), log=TRUE))}
> obj <- MakeADFun(l,par)
> fit<-nlminb(obj$par,obj$fn,obj$gr)
> abline(fit$par[2:1], lwd=3, col="orange")
> exp(fit$par[3])                                  # sigma ca. 0.21
> arrows(x,y,x,fit$par[1]*x+fit$par[2], code=0)
```

Exercise 2.1: non-linear regression

- A model which is often used to describe growth of fish is the von Bertalanffy growth function:

$$L(a) = L_\infty(1 - e^{-ka})$$

- $L(a)$ is length at age a.
- The two model parameters are the asymptotic length $L_\infty > 0$ and the growth rate $k > 0$
- The data set `length.tab` contains corresponding measurements of age and length of 100 fish of the species 'opaleye'
- In this exercise we wish to estimate the model parameters in the model:

$$\log(L_i) = \log(L_\infty) + \log(1 - e^{-ka_i}) + \varepsilon_i \ , \varepsilon \sim \mathbf{N}(0, \sigma^2) \text{ independently } \ i = 1 \dots 100$$

- Make also a figure showing the observations and the predicted curve

Exercise 2.2: Thinking about likelihood (from Pawitan)

The following shows the heart rate (beats/minute) of a person, measured throughout the day:

```
73, 75, 84, 76, 93, 79, 85, 80, 76, 78, 80
```

Assume the data are an iid sample from $\mathcal{N}(\mu, \sigma^2)$, where σ^2 is known as the observed sample variance. Denote the ordered values by $x_{(1)}, x_{(2)}, \ldots, x_{(11)}$. Draw and compare the likelihood of μ if:

a) The whole data $x_1, x_2, \ldots, x_{11}$ are reported

b) only the sample mean $\bar{x}$ is reported

c) only the sample median $x_{(6)}$ is reported

d) only the minimum $x_{(1)}$ and maximum $x_{(11)}$ are reported

e) only the two lowest values $x_{(1)}$ and $x_{(2)}$ are reported

Maximum likelihood estimator and Hessian — in general

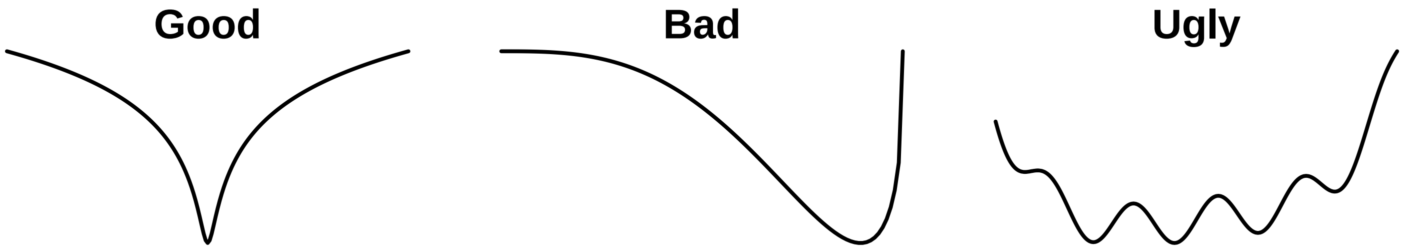
- A sensible estimate of the model parameters is to choose the values that maximize the likelihood for the actual observations.

$$\widehat{\theta} = \underset{\theta}{\text{argmin}}\, \ell(y|\theta)$$

- The curvature of the negative log likelihood function gives an asymptotic estimate of the variance of the maximum likelihood estimator:

$$\widehat{\text{var}(\widehat{\theta})} = \left(\left. \frac{\partial^2 \ell(y|\theta)}{\partial \theta^2} \right|_{\theta=\widehat{\theta}} \right)^{-1}$$

- The matrix $\mathcal{H}(\widehat{\theta}) = \left(\left. \frac{\partial^2 \ell(y|\theta)}{\partial \theta^2} \right|_{\theta=\widehat{\theta}} \right)$ is often referred to as “the hessian matrix”
- Asymptotically we know that $\widehat{\theta} \sim \mathcal{N}(\theta, \mathcal{H}(\theta)^{-1})$, but in practice we may be far from the asymptotic behaviour.



Choosing parameterization

- Consider the model:

$$X \sim \mathsf{Bin}(100, p)$$

- Let's say we have observed $X = 2$

- Want to estimate our model parameter p

```
library(RTMB)
dat <- list(X=2)
par <- list(p=.5)

f<-function(par){-dbinom(dat$X,100,par$p,log=TRUE)}

obj <- MakeADFun(f, par, silent=TRUE)
opt <- nlminb(obj$par, obj$fn, obj$gr, lower=c(0), upper=c(1))
summary(sdreport(obj))

#   Estimate Std. Error
# p     0.02 0.01398284
```

files/p1.R

- See the problem?

Simple bounds on a parameter via transformation

- Consider same model and observation, but now parametrized as:

$$X \sim \text{Bin}(100, p), \quad \text{where } \text{logit}(p) = \alpha$$

- Now we write as:

```
library(RTMB)
dat <- list(X=2)
par <- list(alpha=0)

f<-function(par){
  p <- plogis(par$alpha)                                  ### Notice exp(alpha)/(1+exp(alpha))
  -dbinom(dat$X,100,p,log=TRUE)
}

obj <- MakeADFun(f, par, silent=TRUE)
opt <- nlminb(obj$par, obj$fn, obj$gr)
sdr<-sdreport(obj)
summary(sdr)
#       Estimate Std. Error
# alpha -3.89182  0.7142857
pl<-as.list(sdr,"Est")
plsd<-as.list(sdr,"Std")
plogis(pl$alpha+c(-2,2)*plsd$alpha)
# 0.004867034 0.078475060                                 ### Use same transformation to calculate CI
```

files/p2.R

Exercise 2.3: Thinking about parameterization

Suggest how to use transformation to parametrize a parameter that is

a) only positive

b) only negative

c) between 2 and 5

d) an increasing vector

Getting results out

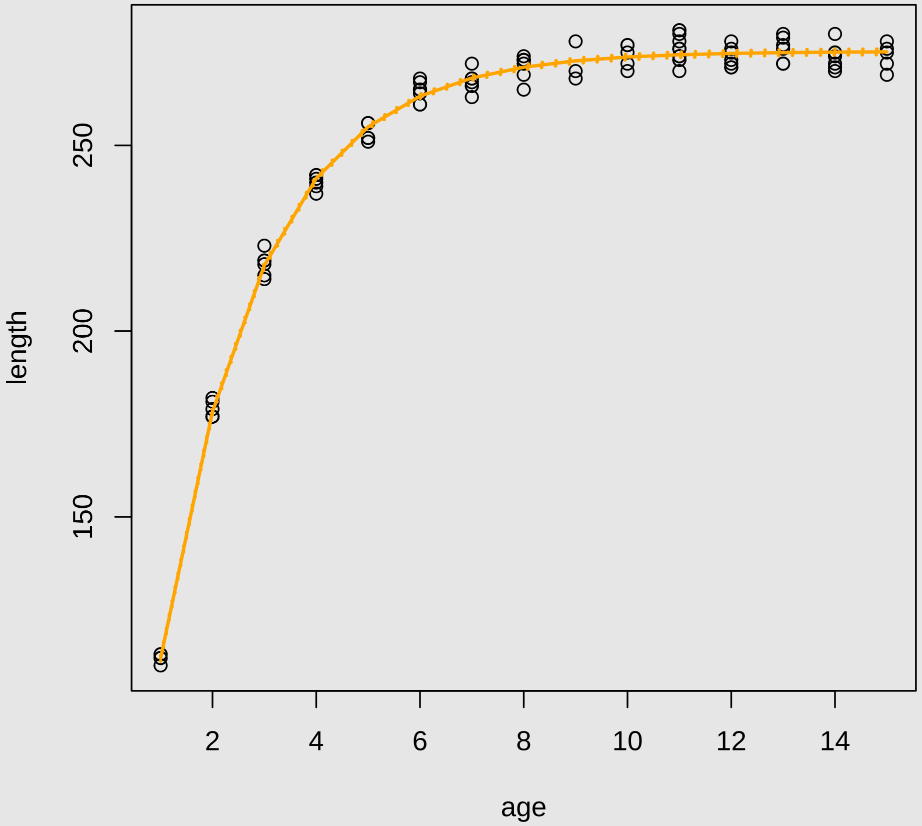
- If estimated standard errors are not needed, then the
  - `REPORT(X)` in the negative log likelihood function
  - `obj$report()$X` after optimization
- If estimated standard errors are needed, then use
  - `ADREPORT(X)` in the negative log likelihood function (for derived quantities)
  - `summary(sdreport(obj))` after optimization
- To get estimates and standard deviations in the same format as they entered the parameter list, try:
  - Parameter list: `pl <- as.list(sdreport(obj),"Est")`
  - Parameter Sd list: `plsd <- as.list(sdreport(obj),"Std")`
- And similar for ad-reported items:
  - Derived estimates: `plr <- as.list(sdreport(obj), what="Est", report=TRUE)`
  - Sd corresponding: `plrsd <- as.list(sdreport(obj), what="Std", report=TRUE)`

Getting results out — example

```
library(RTMB)
par <- list(logLinf=0, logK=0, logSigma=0)
dat <- read.table("files/length.tab", header=TRUE)
l <- function(par){
  getAll(par, dat)
  Linf <- exp(logLinf)
  k <- exp(logK)
  sigma <- exp(logSigma)
  pred <-  log(Linf) + log(1-exp(-k*age))
  ADREPORT(pred)
  -sum(dnorm(log(length),pred,sd=sigma,log=TRUE))
}
obj <- MakeADFun(l,par, silent=TRUE)
fit <- nlminb(obj$par,obj$fn,obj$gr)

sdr <- sdreport(obj)
pl <- as.list(sdr, "Est")
plsd <- as.list(sdr, "Std")
plr <- as.list(sdr, "Est", report=TRUE)
plrsd <- as.list(sdr, "Std", report=TRUE)

plot(dat)
o<-order(dat$age)
lines(dat$age[o],exp(plr$pred[o]), lwd=2, col="orange")
lines(dat$age[o],exp(plr$pred[o]-2*plrsd$pred[o]), lwd=2, col="orange", lty="dotted")
lines(dat$age[o],exp(plr$pred[o]+2*plrsd$pred[o]), lwd=2, col="orange", lty="dotted")
```



out.R

Likelihood functions from a few known models

Poisson: $x_i \sim Pois(\lambda)$ independent

$$\ell(x|\lambda) = \lambda n - \log(\lambda)\sum x_i + \sum \log(x_i!)$$

```
nll = -sum(dpois(X,lambda,log=TRUE));
```

Normal: $x_i \sim \mathcal{N}(\mu, \sigma^2)$ independent

$$\ell(x|\mu,\sigma^2) = \frac{n}{2}\log(2\pi\sigma^2) + \frac{1}{2\sigma^2}\sum(x_i - \mu)^2$$

```
nll= -sum(dnorm(X,mu,sigma,log=TRUE));
```

Binomial: $x_i \sim Bin(N_i, p)$ independent (assume N_i known)

$$\ell(x|p) = -\sum \log\binom{N_i}{x_i} - \log(p)\sum x_i - \log(1-p)\sum(N_i - x_i)$$

```
nll = -sum(dbinom(X,N,p,log=TRUE));
```

Notation: In the above `lambda`, `mu`, `sigma`, and `p` are model parameters, `X` is the observation vector, and `N` is the number of observations, except for the binomial where `N` is a vector of the number of trials.

Asymptotic results

- A frequent starting point for asymptotic results is to approximate $\ell(\theta) = -\log L(\theta)$ with its 2. order Taylor approximation:

$$\ell(\theta) \approx \ell(\hat{\theta}) + \ell'(\hat{\theta})(\theta - \hat{\theta}) + \frac{1}{2}\ell''(\hat{\theta})(\theta - \hat{\theta})^2$$

- Or the multivariate version:

$$\ell(\theta) \approx \ell(\hat{\theta}) + (\theta - \hat{\theta})^T \ell'(\hat{\theta}) + \frac{1}{2}(\theta - \hat{\theta})^T \ell''(\hat{\theta})(\theta - \hat{\theta})$$

- Remember that:
  - $\ell'(\hat{\theta}) = 0$, because that is how we find $\hat{\theta}$.
  - The information is defined as $I(\hat{\theta}) = \ell''(\hat{\theta})$
  - Asymptotically var$(\hat{\theta}) = I(\hat{\theta})^{-1}$
  - If $x \sim N(0,1)$, then x^2 followers a χ^2_1-distribution
- Asymptotically $\hat{\theta}$ followers a normal distribution $N(\theta, I(\theta)^{-1})$

Likelihood ratio test for a single parameter

- Consider the situation where we have model (M_1) and are interested in the hypothesis

  $$H_0 : \theta = \theta_0$$

  for a single model parameter.
- The model M_0 where θ is restricted to be equal to θ_0 is called a sub-model, because model M_0 is a special case under model M_1.
- If we optimize each model we get two estimates $\hat{\theta}_1$ and $\hat{\theta}_0$
- Consider the ratio of the likelihoods:

  $$Q_{M_1 \to M_0} = \frac{L(\hat{\theta}_0)}{L(\hat{\theta}_1)}$$
- This number is between 0 and 1 (why?).
- If 'near' 1 it means that H_0 is acceptable (why?).
- If 'near' 0 it means that H_0 not acceptable, as model M_1 is describing the data much better than model M_0.

- But how close to 0 is random?
- Start by assuming H_0 is true.
- If we look at:

$$G_{M_1 \to M_0} = -2 \log Q_{M_1 \to M_0} = -2 \log \frac{L(\hat{\theta}_0)}{L(\hat{\theta}_1)}$$

- Then the asymptotic distribution of this quantity is:

$$G_{M_1 \to M_0} \sim \chi^2_1$$

  because it is $I(\hat{\theta}_0)(\hat{\theta}_1 - \hat{\theta}_0)^2$, which is a standardized normal squared (remember?)

- So we can calculate the p-value by:

$$P_{M_1 \to M_0} = P\left(\chi^2_1 \geq G_{M_1 \to M_0}\right)$$

- If this is small (often defined as $< 5\%$) the actual observations matches M_0 poorly and the model reduction is rejected.

Likelihood ratio test - general case

- Assume model M_0 is a sub model of model M_1 (this is for instance the case if a free model parameter in M_1 is set to a fixed value in M_0)
- We can calculate the test statistic $G_{M_1 \to M_0}$ for reducing model M_1 to model M_0 by:

$$G_{M_1 \to M_0} = 2(\ell(y|\widehat{\theta}_0) - \ell(y|\widehat{\theta}_1))$$

- If the two optimal fits are "almost equal" the model reduction is accepted, if the fits are very different the model reduction is rejected
- Asymptotically G followers a χ^2–distribution, so the P-value is given by:

$$P_{M_1 \to M_0} = P\left(\chi^2_{\mathsf{dim}(M_1) - \mathsf{dim}(M_0)} \geq G_{M_1 \to M_0}\right)$$

- If this is small (often defined as $< 5\%$) the actual observations matches B poorly and the model reduction is rejected.

Exercise 2.4: Test separate growth for male and female

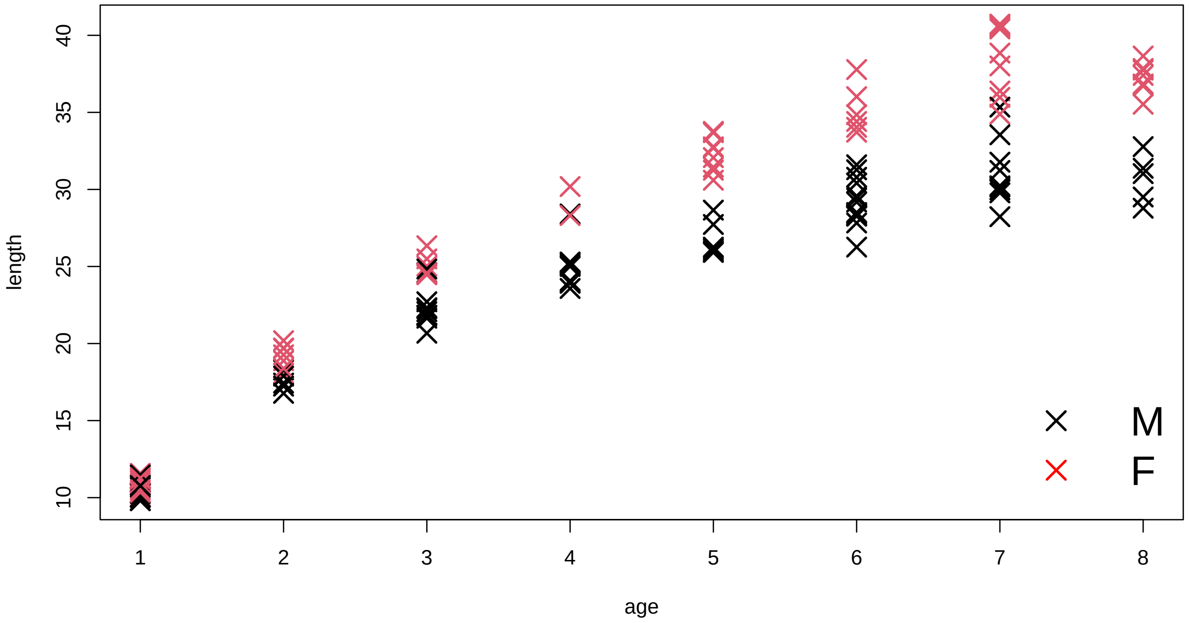
- Consider again the data in the file `length2.tab` and used the model you set up for this data.

- Compute the likelihood-ratio tests for:

  H0: $k_{\text{male}} = k_{\text{female}}$

  H0: $L_{\infty,\text{male}} = L_{\infty,\text{female}}$

  H0: same growth pattern for male and female.



Maximum likelihood estimator, asymptotic distribution, and test

- Parameters are estimateed as those that maximize likelihood for the observations, so with $\ell(\theta) = -\log L(y|\theta)$ as:

$$\widehat{\theta} = \underset{\theta}{\text{argmin}}\, \ell(y|\theta)$$

- Asymptotically:

$$\widehat{\theta} \sim \mathcal{N}(\theta, E(\mathcal{H}(\theta))^{-1}), \quad \text{where} \quad \mathcal{H}(\theta) = \left(\frac{\partial^2 \ell(y|\theta)}{\partial \theta^2}\right)$$

- In practice we plug in $\widehat{\theta}$ and $\mathcal{H}(\widehat{\theta})^{-1} = \left(\frac{\partial^2 \ell(y|\theta)}{\partial \theta^2}\Big|_{\theta=\widehat{\theta}}\right)^{-1}$ as mean vector and covariance matrix respectively
- The likelihood ratio test statistic for H0, defining a sub-model M_0 of the general model M_1, is defined as:

$$G_{M_1 \to M_0} = -2\log\left(\frac{L(\widehat{\theta}_0)}{L(\widehat{\theta}_1)}\right) = 2(\ell(y|\widehat{\theta}_0) - \ell(y|\widehat{\theta}_1))$$

- Asymptotically G followers a χ^2–distribution, so the P-value is given by:

$$P_{M_1 \to M_0} = P\left(\chi^2_{\text{dim}(M_1) - \text{dim}(M_0)} \geq G_{M_1 \to M_0}\right)$$

- If this is small (often defined as $< 5\%$) the observations matches H_0 poorly and the model reduction is rejected.

Collapsing parameters, or fixing them

- The `map` argument of the `MakeADFun` can be used to couple elements in a parameter object
- If we have a parameter vector `alpha` of length 4, then the statement:

```
obj <- MakeADFun(f, par, map=list(alpha=factor(c(1,2,3,3))))
```

- will collapse the last two parameters.
- They will be initialized to the mean of the last two initializations
- The optimizer will estimate a common value for both parameters
- This structure is perfect for testing many model hypotheses
- In addition if `NA` is set, as in:

```
obj <- MakeADFun(f, par, map=list(alpha=factor(c(1,2,NA,4))))
```

- then the optimizer treat that parameter (here the third) as fixed.

Exercise: Use of the map argument

- Consider the data set `InsectSprays`, which is available in R
- We will use the model: $\text{count}_i \sim \text{Pois}(\lambda_i)$, where $\log \lambda_i = \alpha(\text{spray}_i)$
- This can be implemented as:

```r
library(RTMB)

# for data we use the built-in data "InsectSprays"
par <- list(logAlpha=rep(0,nlevels(InsectSprays$spray)))
f<-function(par){
  getAll(InsectSprays, par)
  nll <- 0
  for(i in 1:length(count)){
    lambda <- exp(logAlpha[spray[i]])
    nll <- nll - dpois(count[i],lambda,log=TRUE)
  }
  nll
}
obj <- MakeADFun(f, par)
opt <- nlminb(obj$par, obj$fn, obj$gr)
```

files/insect.R

- Use the `map` argument to test the hypothesis that spray $\alpha(A) = \alpha(B) = \alpha(F)$
- Can the mean count for the spray ‘A’, ‘B’ and ‘F’ and be assumed to be equal to 15.

(try to test these hypothesis without modifying the funtion ‘f’)