Fishery Population Analysis

Toshihide KITAKADO

Lecture 12 Regression analyses

Regression models

	Distributional assumption	Regression component	R function
Normal linear model	Normal	Linear	“lm”
Normal nonlinear model	Normal	Nonlinear	“nls”
Generalized linear model (GLM)	Exponential family (Normal, Gamma, Binomial, Poisson etc)	Linear through “a link function”	“glm”
Additive model	Normal	Nonparametric	“gam”
Generalized additive model (GAM)	Exponential family (Normal, Gamma, Binomial, Poisson etc)	Nonparametric through “a link function”	“gam”

Least square method for simple linear regression

Linear relationship $y = \alpha + \beta x$

Observation $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$

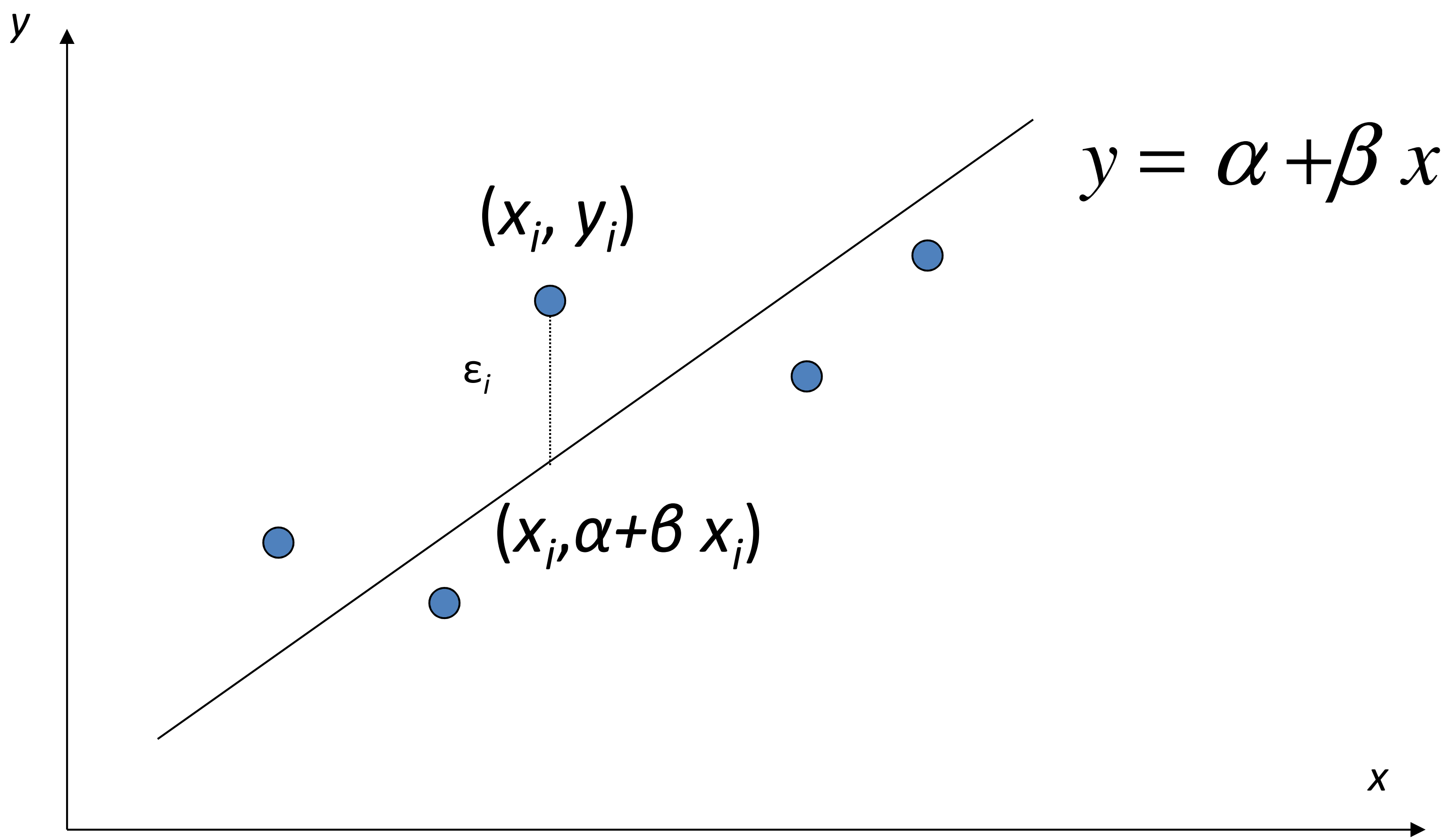
Simple linear regression

$$y_i = \alpha + \beta x_i + \varepsilon_i \;, \quad \varepsilon_i \sim N(0, \sigma^2)$$

Least square estimation for α, β

$$S(\alpha, \beta) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2 \rightarrow \min$$

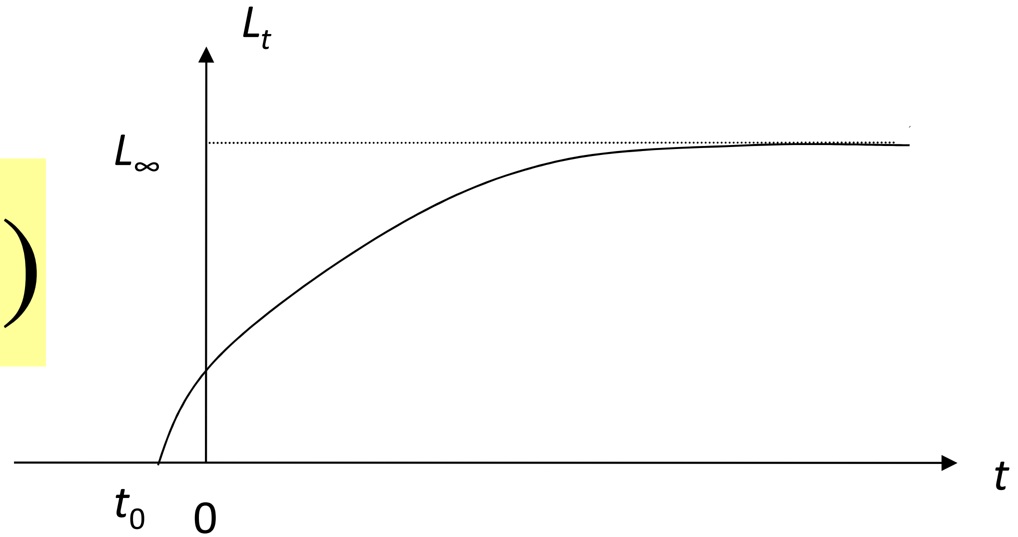
Least square method for simple linear regression



$$S(\alpha, \beta) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2 \rightarrow \min$$

von Bertalanffy formula

$$L_t = L_\infty (1 - e^{-k(t - t_0)})$$



t : age

L_t :Length at the age t

L_∞ :Asymptotic length

k : Growth coefficient

t_0 : age at which L_t=0 satisfies

Rainbow trout allometry

```
#Reading data
Data <- read.csv("rainbowtrout.csv", header=T)
names(Data)
Length <- Data$Length
Weight <- Data$Weight

#Regression for logarithms of data
plot(log(Length), log(Weight))
res <- lm(log(Weight)~log(Length))
summary(res)
abline(res)

par(mfrow=c(2,2))
plot(res)
Est <- coef(res)
CI <- confint(res)
```

Non-linear regression

Nonlinear relationship $y = f(x;\theta)$

Observation $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$

Model (additive error structure)

$$y_i = f(x_i;\theta) + \varepsilon_i \ , \quad \varepsilon_i \sim N(0, \sigma^2)$$

Least square method

$$S(\theta) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - f(x_i;\theta))^2 \rightarrow \min$$

Non-linear regression

$$y_i = L(t_i\,|\,\theta) + \varepsilon_i\,, \qquad \varepsilon_i \sim N(0,\sigma^2)$$

$$y_i \sim N(L(t_i\,|\,\theta),\sigma^2)$$

$$L(\theta,\sigma^2) = \prod_{i=1}^{n} f(y_i\,|L(t_i\,|\,\theta),\sigma^2)$$

$$l(\theta,\sigma^2) = \log L(\theta,\sigma^2) = \sum_{i=1}^{n} \log f(y_i\,|\,L(t_i\,|\,\theta),\sigma^2)$$

$$= \sum_{i=1}^{n} \log \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y_i - L(t_i|\theta))^2}$$

$$= -\frac{n}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2}\boxed{\sum_{i=1}^{n}(y_i - L(t_i\,|\,\theta))^2}$$

Minimize this term wrt θ

$$\hat{\theta} = (\hat{L}_\infty, \hat{K}, \hat{t}_0)$$

In case of von Bertalanffy

$$L(t\,|\,\theta) = L_\infty\,(1 - e^{-K(t-t_0)})$$

$$\theta = (L_\infty, K, t_0)$$

$$\frac{\partial}{\partial\sigma^2} l(\theta,\sigma^2) = \frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(y_i - L(t_i\,|\,\theta))^2 = 0$$

$$\Rightarrow\ \sigma^2(L_\infty,\, K,\, t_0) = \frac{1}{n}\sum_{i=1}^{n}(y_i - L(t_i\,|\,\theta))^2$$

$$\Rightarrow\ \hat{\sigma}^2 = \sigma^2(\hat{L}_\infty,\, \hat{K},\, \hat{t}_0) = \frac{1}{n}\sum_{i=1}^{n}(y_i - L(t_i\,|\,\hat{\theta}))^2$$

Estimate of error variance

Estimation of growth curve (I)

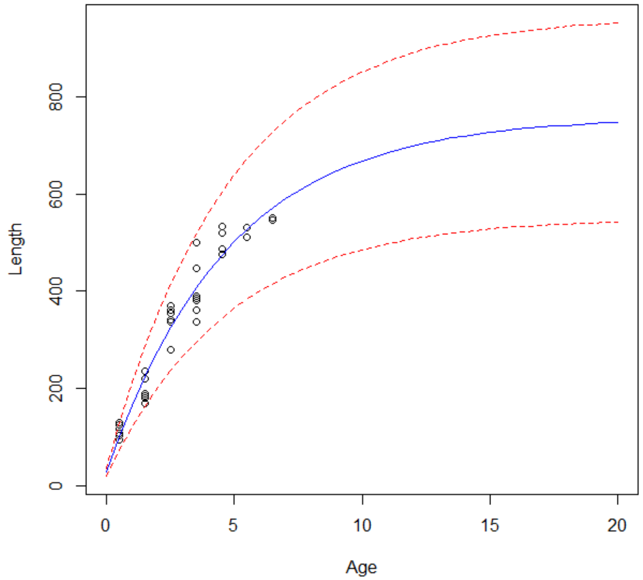
```
#Reading the data
Growthdata<-read.csv("growthdata.csv", header=T)
attach(Growthdata)
plot(Age, Length)

#Non-linear regression by the least square method
library(stats)
Start<-c(L=600,K=0.1,t0=0)
res.nls<-nls(Length~L*(1-exp(-K*(Age-t0))),   start=Start)
summary(res.nls)
coef(res.nls)
confint(res.nls, level = 0.95)
```

Estimation of growth curve (II)

```
newx<-seq(0,20, 0.5)
pred<-predict(res.nls, list(Age=newx), int="c")
#BUT…予測値の信頼区間が自動的に出るようには未だなってない!!
#したがってデルタ法で自ら計算
av<-vcov(res.nls)
L<-as.numeric(coef(res.nls))[1]
K<-as.numeric(coef(res.nls))[2]
t0<-as.numeric(coef(res.nls))[3]
dL<-function(t){1-exp(-K*(t-t0))}
dK<-function(t){(t-t0)*exp(-K*(t-t0))}
dt0<-function(t){K* exp(-K*(t-t0))}
D<-array(0,c(length(newx), 3))
D[,1]<-sapply(newx, dL)
D[,2]<-sapply(newx, dK)
D[,3]<-sapply(newx, dt0)
av.pred<-D %*% av %*% t(D)   #delta method
```

Estimation of growth curve (III)




```
#continued
av.pred<-D %*% av %*% t(D)
se.pred<-sqrt(diag(av.pred))
c<-qnorm(0.975,0,1)
Lower <- pred - c*se.pred
Upper <- pred + c*se.pred
pcon<-cbind(pred, Lower, Upper)

plot(Age, Length, xlim=c(0,20), ylim=range(pcon))
matlines(newx,pcon, lty=c(1,2,2), col=c("blue","red","red"))

AIC(res.nls)

AIC(lm(Length~Age))
```

Notes

Additive model (sd: constant)

$$y_i = L(t_i \mid \theta) + \varepsilon_i$$

```
res.nls<-nls(Length~L*(1-exp(-K*(Age-t0))),   start=Start)
AIC(res.nls)
```

Multiplicative model (sd: proportional to length)

$$\log y_i = \log L(t_i \mid \theta) + \varepsilon_i$$

$$y_i = L(t_i \mid \theta)\; e^{\varepsilon_i}$$

```
res.nls2<-nls(log(Length)~log(L*(1-exp(-K*(Age-t0)))),   start=Start)
AIC(res.nls2)
 #not comparable to model 1 because the data are different
AIC(res.nls2) + 2*sum(log(Length))  # comparable
```

Rev: Estimation of a binomial parameter

Binomial distribution

$$Y_i \sim Bin(N_i, p) \quad (i = 1, \ldots, n)$$

$$\Pr(Y_i = y_i) = \binom{N_i}{y_i} p^{y_i} (1-p)^{N_i - y_i}$$

The likelihood function

$$L(p) = \prod_{i=1}^{n} \binom{N_i}{y_i} p^{y_i} (1-p)^{N_i - y_i}$$

The log-likelihood function

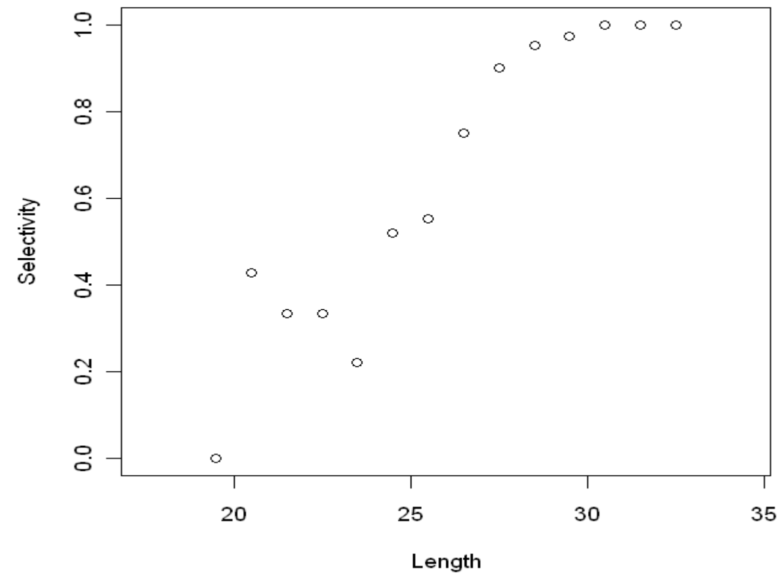
$$l(p) = \log L(p) = \sum_{i=1}^{n} \log \binom{N_i}{y_i} + \sum_{i=1}^{n} [y_i \log p + (N_i - y_i) \log(1-p)]$$

Rev: Estimation of binomial parameters

Binomial distribution

$$Y_i \sim Bin(N_i, p_i) \quad (i = 1, \ldots, n)$$

$$\Pr(Y_i = y_i) = \binom{N_i}{y_i} p_i^{y_i} (1 - p_i)^{N_i - y_i}$$



The likelihood function

$$L(p_1, \ldots, p_n) = \prod_{i=1}^{n} \binom{N_i}{y_i} p_i^{y_i} (1 - p_i)^{N_i - y_i}$$

The log-likelihood function

$$l(p_1, \ldots, p_n) = \log L(p_1, \ldots, p_n) = \sum_{i=1}^{n} \log \binom{N_i}{y_i} + \sum_{i=1}^{n} [y_i \log p_i + (N_i - y_i) \log(1 - p_i)]$$

Estimation of regression coefficients in a logistic curve

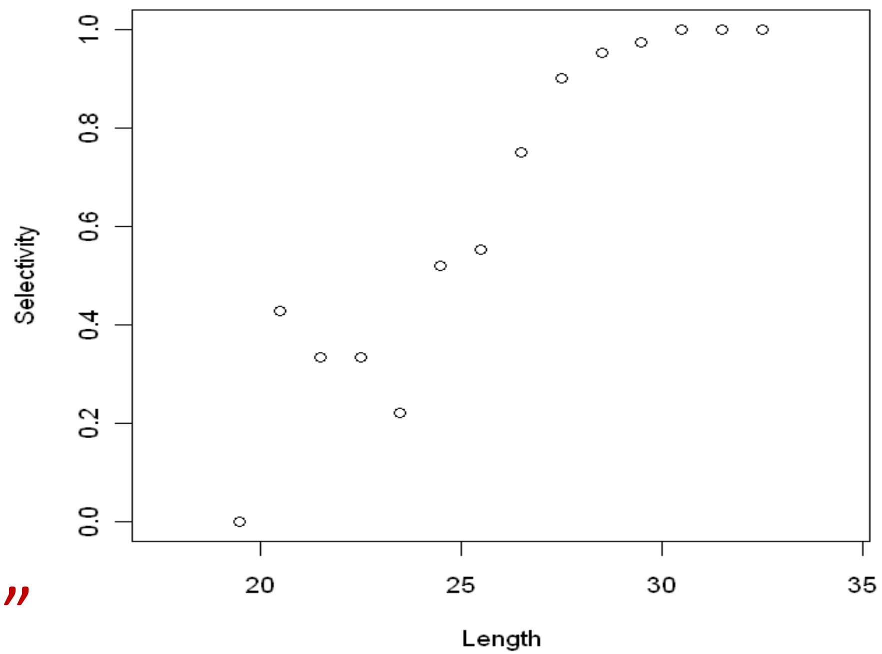
Binomial distribution

$$Y_i \sim Bin(N_i, s(l_i)) \quad (i = 1, \ldots, n)$$

$$\Pr(Y_i = y_i) = \binom{N_i}{y_i} s(l_i)^{y_i} (1 - s(l_i))^{N_i - y_i}$$

$$s(l) = \frac{e^{a+bl}}{1 + e^{a+bl}}$$

$$\log \frac{s(l)}{1 - s(l)} = a + bl$$



This is called a "link function"

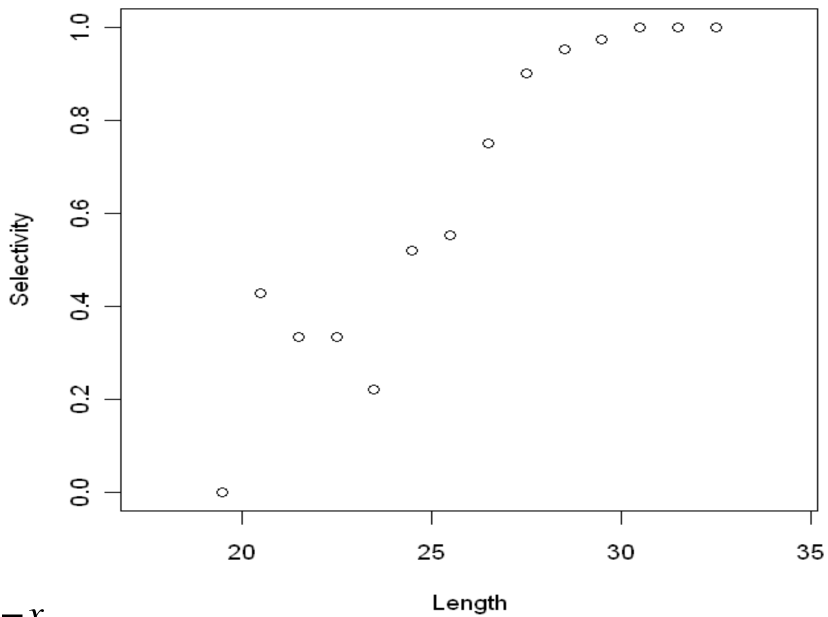
Estimation of regression coefficients in a logistic curve

Binomial distribution

$$X_i \sim Bin(N_i, s(l_i)) \quad (i = 1, \ldots, n)$$

$$\Pr(X_i = x_i) = \binom{N_i}{x_i} s(l_i)^{x_i} (1 - s(l_i))^{N_i - x_i}$$

$$s(l) = \frac{e^{a+bl}}{1 + e^{a+bl}}$$



The likelihood function

$$L(a,b) = \prod_{i=1}^{n} \binom{N_i}{x_i} s(l_i)^{x_i} (1 - s(l_i))^{N_i - x_i}$$

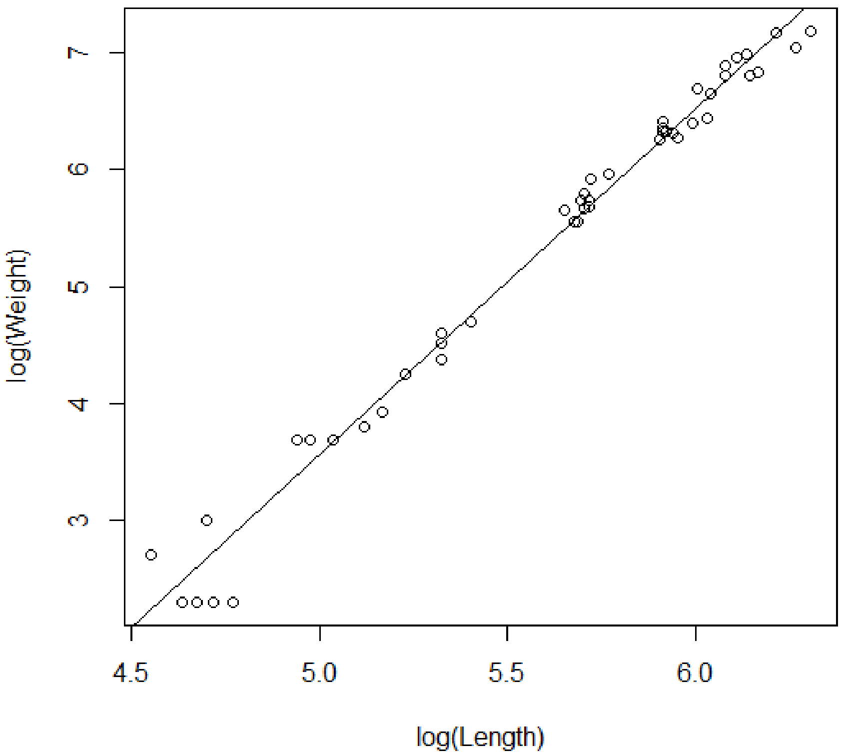
The log-likelihood function

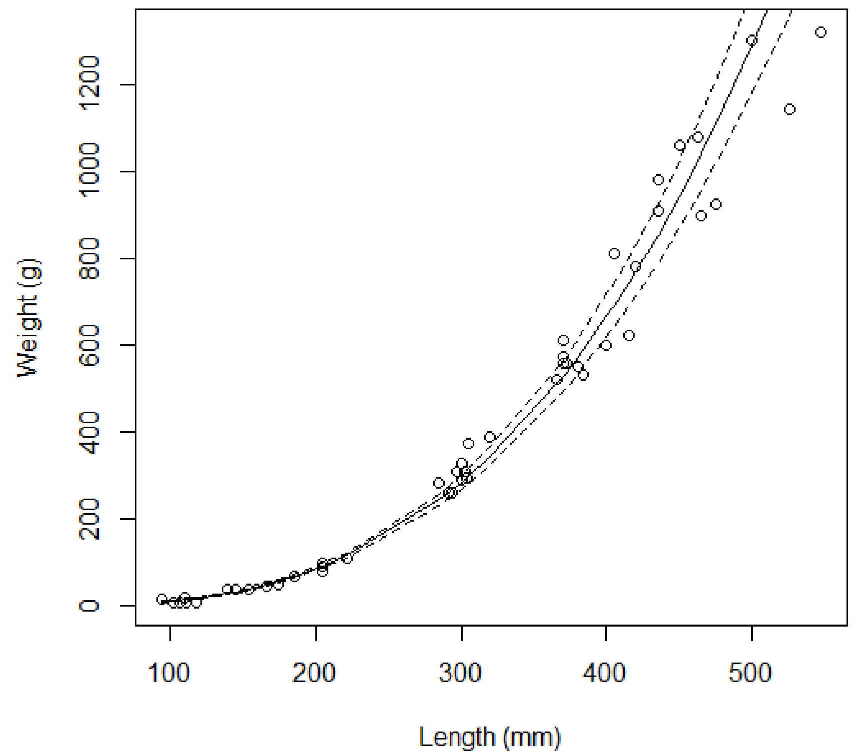
$$l(a,b) = \log L(a,b) = \sum_{i=1}^{n} \log \binom{N_i}{x_i} + \sum_{i=1}^{n} \left[x_i \log s(l_i) + (N_i - x_i) \log(1 - s(l_i)) \right]$$

Regression models

Normal linear model

$$y_i = \alpha + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2)$$



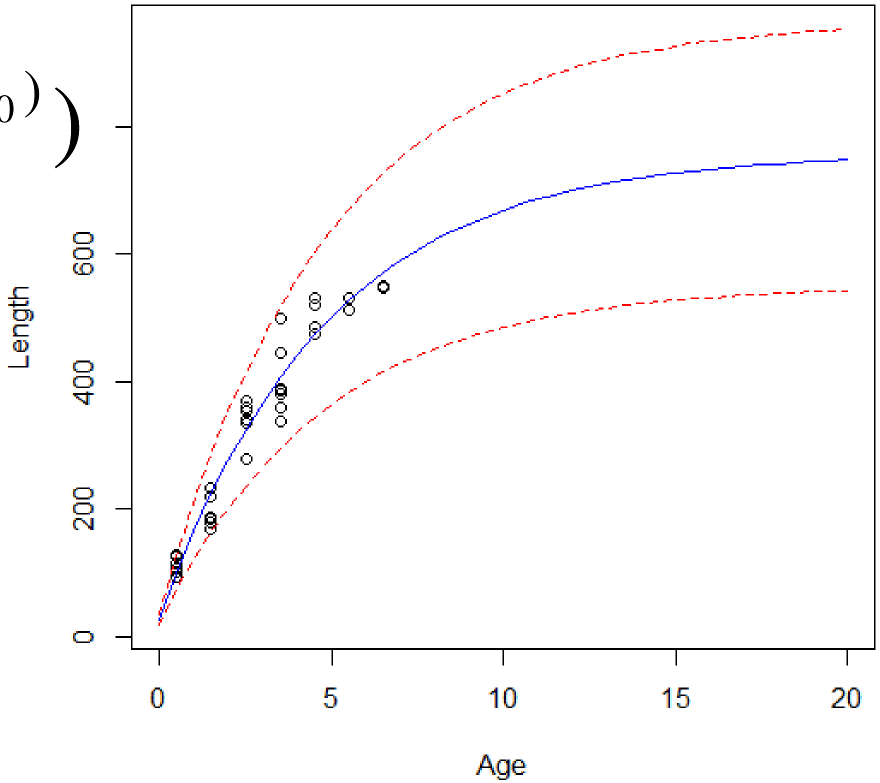


Regression models

Normal non-linear model

$$y_i = f(x_{i1},\ldots,x_{ip};\alpha,\beta_1,\ldots,\beta_p) + \varepsilon_i, \quad \varepsilon_i \sim N(0,\sigma^2)$$

$$f(t;\theta) = L_\infty (1 - e^{-k(t-t_0)})$$



Regression models

Generalized linear model (GLM)

$$y_i \sim Bin(N_i, p(x_i;\theta))$$

$$\log \frac{p(x_i;\theta)}{1-p(x_i;\theta)} = \alpha + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$$

$$y_i \sim Po(\lambda(x_i;\theta))$$

$$\log \lambda(x_i;\theta) = \alpha + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$$

$$y_i \sim N(\mu(x_i;\theta), \sigma^2)$$

$$\mu(x_i;\theta) = \alpha + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$$

Gamma, Inverse-Gaussian.....

Regression models

Additive model (nonparametric regression

$$y_i = f(x_{i1};\beta_1) + \cdots + f(x_{ip};\beta_p) + \varepsilon_i, \quad \varepsilon_i \sim N(0,\sigma^2)$$

Generalized additive model (GAM)

$$y_i \sim Bin(N_i, p(x_i;\theta))$$

$$\log\frac{p(x_i;\theta)}{1-p(x_i;\theta)} = f(x_{i1};\beta_1) + \cdots + f(x_{ip};\beta_p)$$

$$y_i \sim Po(\lambda(x_i;\theta))$$

$$\log\lambda(x_i;\theta) = f(x_{i1};\beta_1) + \cdots + f(x_{ip};\beta_p)$$

$$y_i \sim N(\mu(x_i;\theta),\sigma^2)$$

$$\mu(x_i;\theta) = f(x_{i1};\beta_1) + \cdots + f(x_{ip};\beta_p)$$

What is the Generalized Linear Model (GLM) ?

	Distributional assumption	Regression component	R function
Normal linear model	Normal	Linear	"lm"
Normal nonlinear model	Normal	Nonlinear	"nls"
Generalized linear model (GLM)	Exponential family (Normal, Gamma, Binomial, Poisson etc)	Linear through "a link function"	"glm"
Additive model	Normal	Nonparametric	"gam"
Generalized additive model (GAM)	Exponential family (Normal, Gamma, Binomial, Poisson etc)	Nonparametric through "a link function"	"gam"