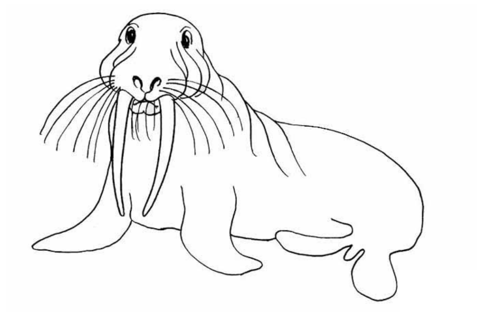
JOURNEES Club WinBUGS

7 décembre 2006

La régression sous WinBugs: une vieille méthode revisitée à partir d'un exemple environnemental.



Eric PARENT, Etienne RIVOT

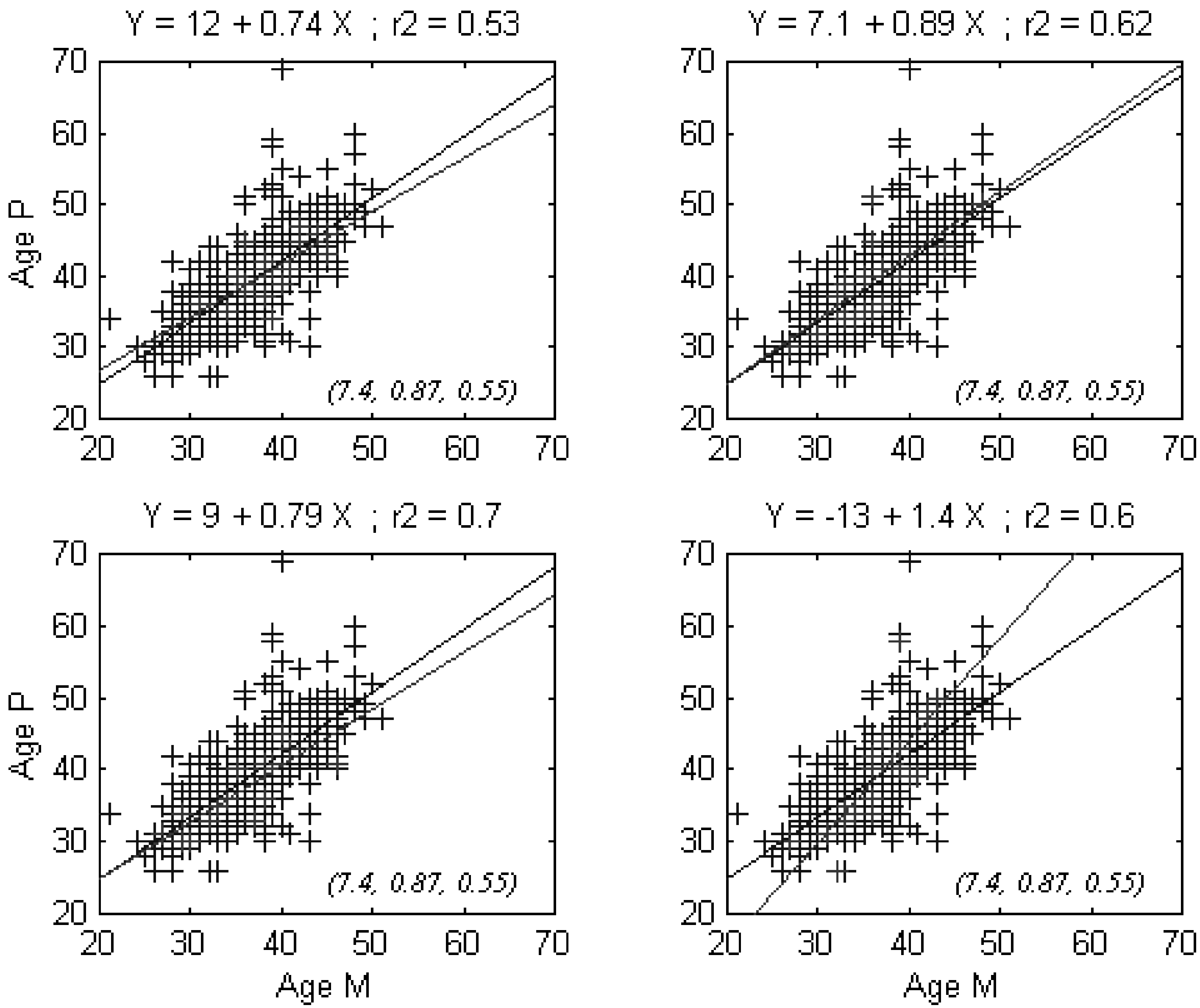
MOdélisation, Risque, Statististique, Environnement

de l 'UMR *MIA 518* INRA/ENGREF/INAPG

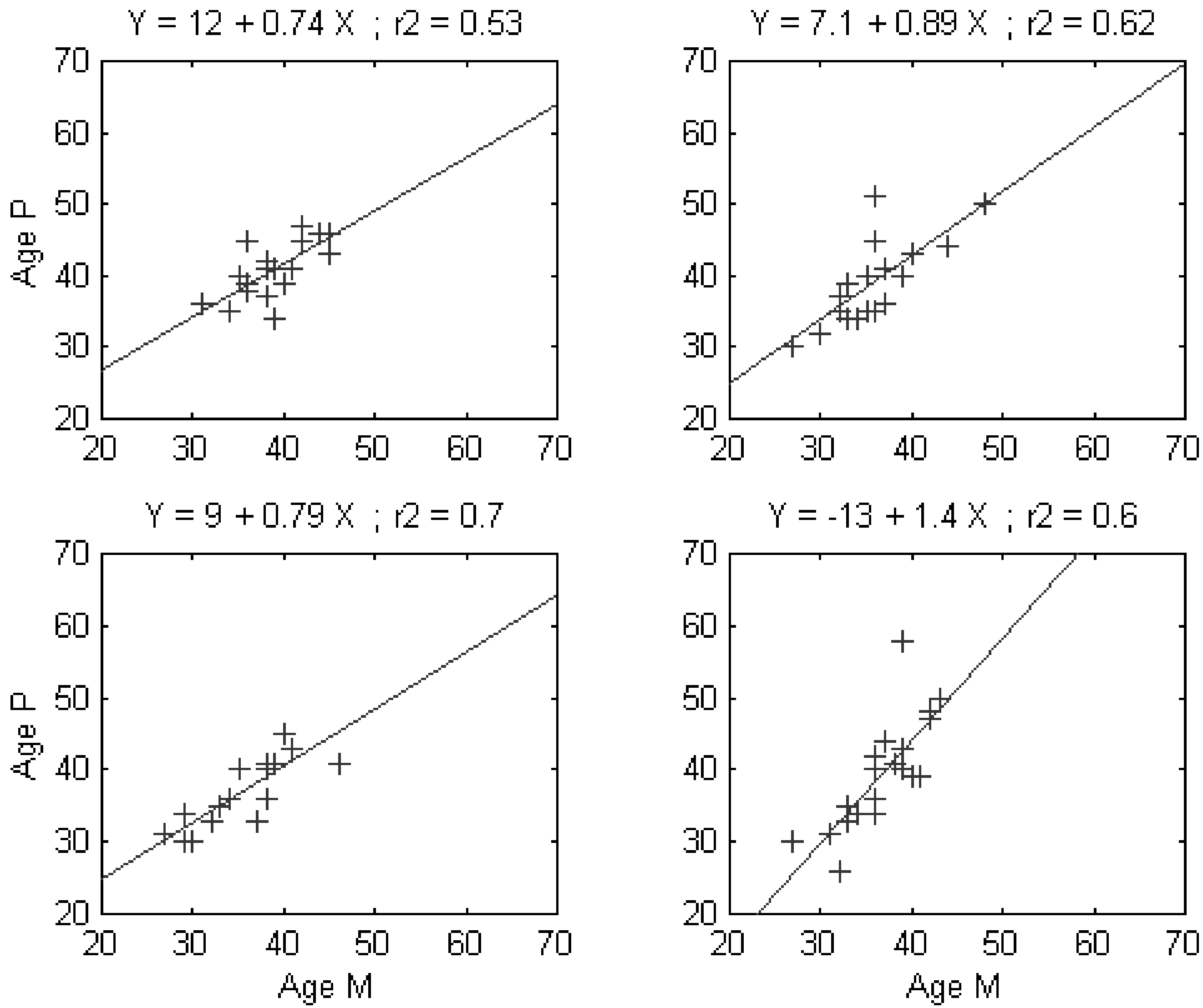
(Math. Info. App. 518)

Basics for regression

- Population vs sample
- True parameters vs estimates
- Hypotheses



Représentation graphique des notions suivantes : population (bleues), modèle (droite bleue), échantillon (croix rouges), modèle ajusté (droite rouge)

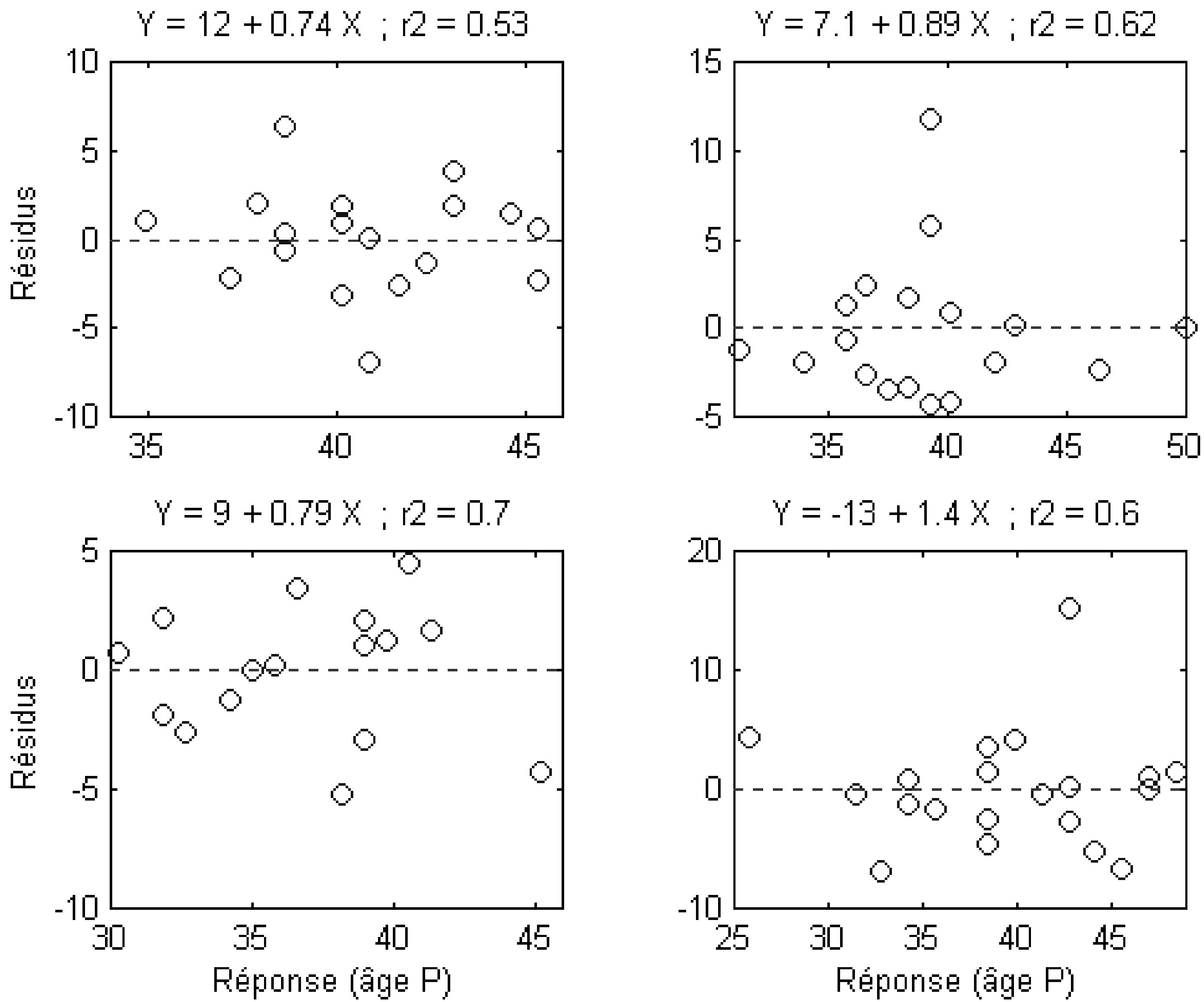


Représentation graphique de l'échantillon (croix rouges) et du modèle ajusté (droite rouge). La droite rouge obtenue est conditionnelle à l'échantillon en main (4 échantillons $\Rightarrow$ 4 droites différentes). Les coefficients (pente, ordonnée à l'origine et r^2) sont donc incertains

Hypotheses

- Residuals are independant
- Residuals are from the same pdf
- Residuals are normal

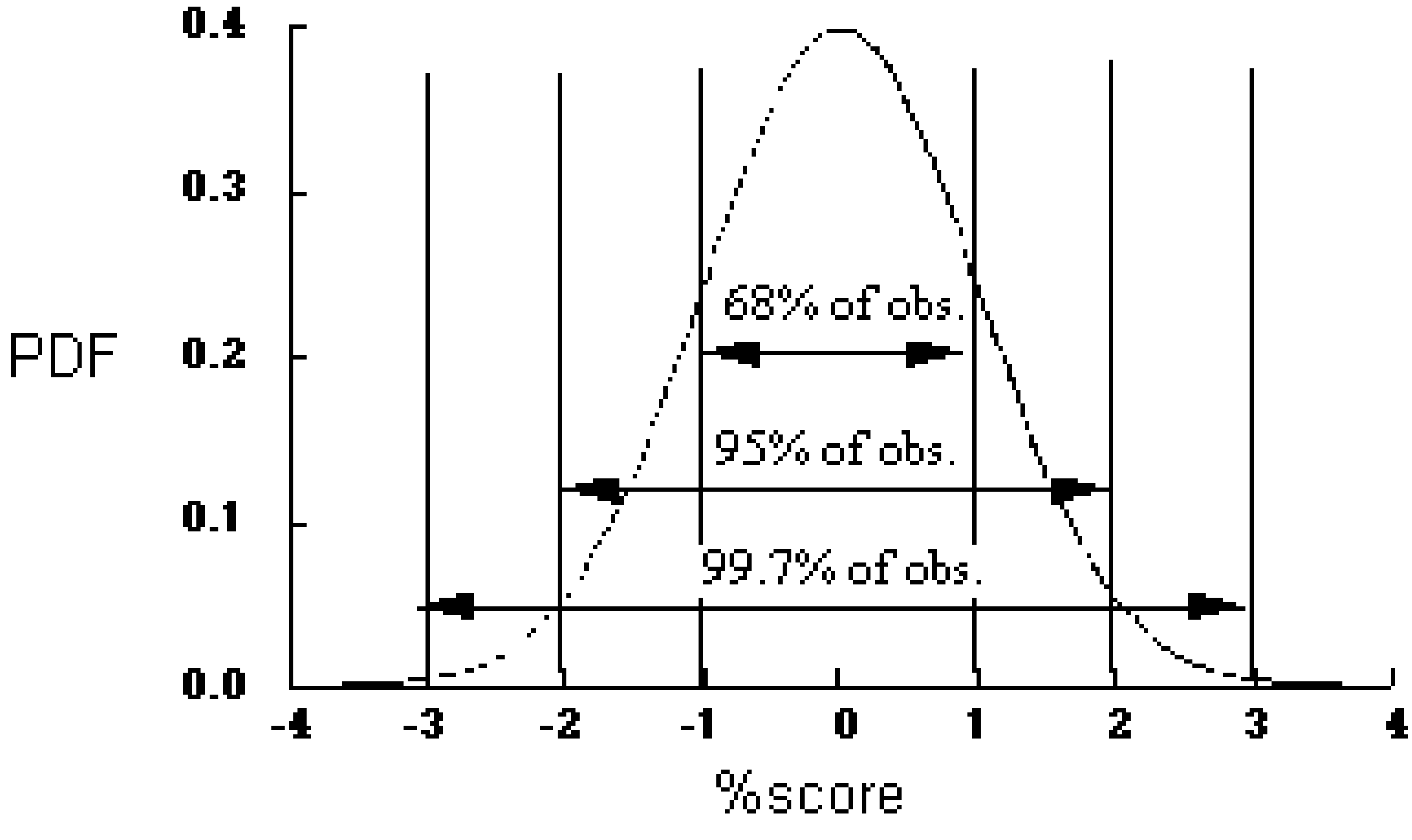
- model structure



Distribution des résidus $e_i = y_i - \hat{y}_i$ en fonction de $\hat{y}_i = a + bx_i$.

$$(\mu, \sigma) known; \quad Z = \frac{X - \mu}{\sigma}$$

Normal PDF



Bayes theory for the linear model

Remember Bayesian statistics ?

prior

likelihood

posterior

$$[\theta|y] = \frac{[\theta]\times[y|\theta]}{[y]}$$

Bayes formula, an updating processor

DATA (y)

Prior information → Posterior information

$[\theta]$ $[\theta|y]$

Bayesian analysis of the linear model

$$y_t = \alpha + \beta x_t + \varepsilon_t \text{ where } \varepsilon_t \sim N(0, \sigma^2)$$

$$[y | \alpha, \beta, \sigma^2, X] = \left(\frac{1}{\sqrt{2\pi}\sigma} \right)^n \times \exp\left(-\frac{1}{2\sigma^2} \{ (Y - X\theta)'(Y - X\theta) \} \right)$$

$$\left[\theta, \frac{1}{\sigma^2} \middle| H \right] \propto \underbrace{\left(\frac{1}{\sigma^2} \right)^{a-1} \exp - \frac{b}{\sigma^2}}_{} \exp\left(\underbrace{-\frac{1}{2}(\theta - \theta_0)' \textstyle\sum_0^{-1} (\theta - \theta_0)}_{} \right)$$

$1/\sigma^2 \sim \text{Gamma}(a,b)$

$\theta \sim N(\theta_0, \Sigma_0)$

$$Y = \begin{Vmatrix} y_1 \\ \ldots \\ y_t \\ \ldots \\ y_n \end{Vmatrix} \quad ; X = \begin{Vmatrix} 1 & x_1 \\ 1 & \ldots \\ \ldots & x_t \\ \ldots & \ldots \\ 1 & x_n \end{Vmatrix} \quad ; \quad \theta = \begin{Vmatrix} \alpha & \beta \end{Vmatrix}$$

Bayesian analysis of the linear model

$$Y = X\theta + \varepsilon \quad \text{where} \quad \varepsilon \sim N(0, \sigma^2 I)$$

$$\hat{\theta} = (X'X)^{-1} X'Y$$ ← Least square estimate

$$[y|\alpha, \beta, \sigma^2, X] = \left(\frac{1}{\sqrt{2\pi\sigma}}\right)^n \times \exp\left(-\frac{1}{2\sigma^2}\left\{(Y - X(\theta - \hat{\theta} + \hat{\theta}))(Y - X(\theta - \hat{\theta} + \hat{\theta}))\right\}\right)$$

Pythagore ' theorem

$$[y|\alpha, \beta, \sigma^2, X] \propto \exp\left(-\frac{1}{2\sigma^2}\left\{(\theta - \hat{\theta})(X'X)(\theta - \hat{\theta}) + (Y - X\hat{\theta})(Y - X\hat{\theta})\right\}\right)$$

θ only!

$$\left[\theta, \frac{1}{\sigma^2}|H\right] \propto \exp\left(-\frac{1}{2}(\theta - \theta_0)'\sum\nolimits_0^{-1}(\theta - \theta_0)\right)\left(\frac{1}{\sigma^2}\right)^{a-1} \exp - \frac{b}{\sigma^2}$$

Bayesian analysis of the linear model

$$Y = X\theta + \varepsilon \quad \text{where} \quad \varepsilon \sim N(0, \sigma^2 I)$$

$$\hat{\theta} = (X'X)^{-1} X'Y \longleftarrow \text{Least square estimate}$$

$$\left[\frac{1}{\sigma^2} \middle| Y, X, \theta\right] \propto \exp -\frac{1}{2}\left((\theta - \theta_0)' \Sigma_0^{-1} (\theta - \theta_0) + \left(\theta - \hat{\theta}\right)\left(\frac{X'X}{\sigma^2}\right)\left(\theta - \hat{\theta}\right) \right)$$

If σ were known, θ would be normal $\boxed{\theta \mid \sigma^2 \sim N(\theta_1, \Sigma_1)}$

$$\begin{cases} \Sigma_1^{-1} = \Sigma_0^{-1} + \left(\dfrac{X'X}{\sigma^2}\right) & \text{Addition of the precision matrices} \\ \Sigma_1^{-1}\theta_1 = \Sigma_0^{-1}\theta_0 + \left(\dfrac{X'X}{\sigma^2}\right)\hat{\theta} & \text{Linear combination between prior mean and least square estimates} \end{cases}$$

Bayesian analysis of the linear model

$$Y = X\theta + \varepsilon \qquad \text{where} \qquad \varepsilon \sim N(0, \sigma^2 I)$$

Gamma (a,b)

$$\left[\frac{1}{\sigma^2} \middle| Y, X, \theta\right] \propto \left(\frac{1}{\sigma^2}\right)^{a-1} \exp -\frac{b}{\sigma^2} \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} \exp -\frac{1}{2\sigma^2}\left((Y - X\theta)'(Y - X\theta)\right)$$

If θ were known, σ^2 would be inverse Gamma (a’,b’)

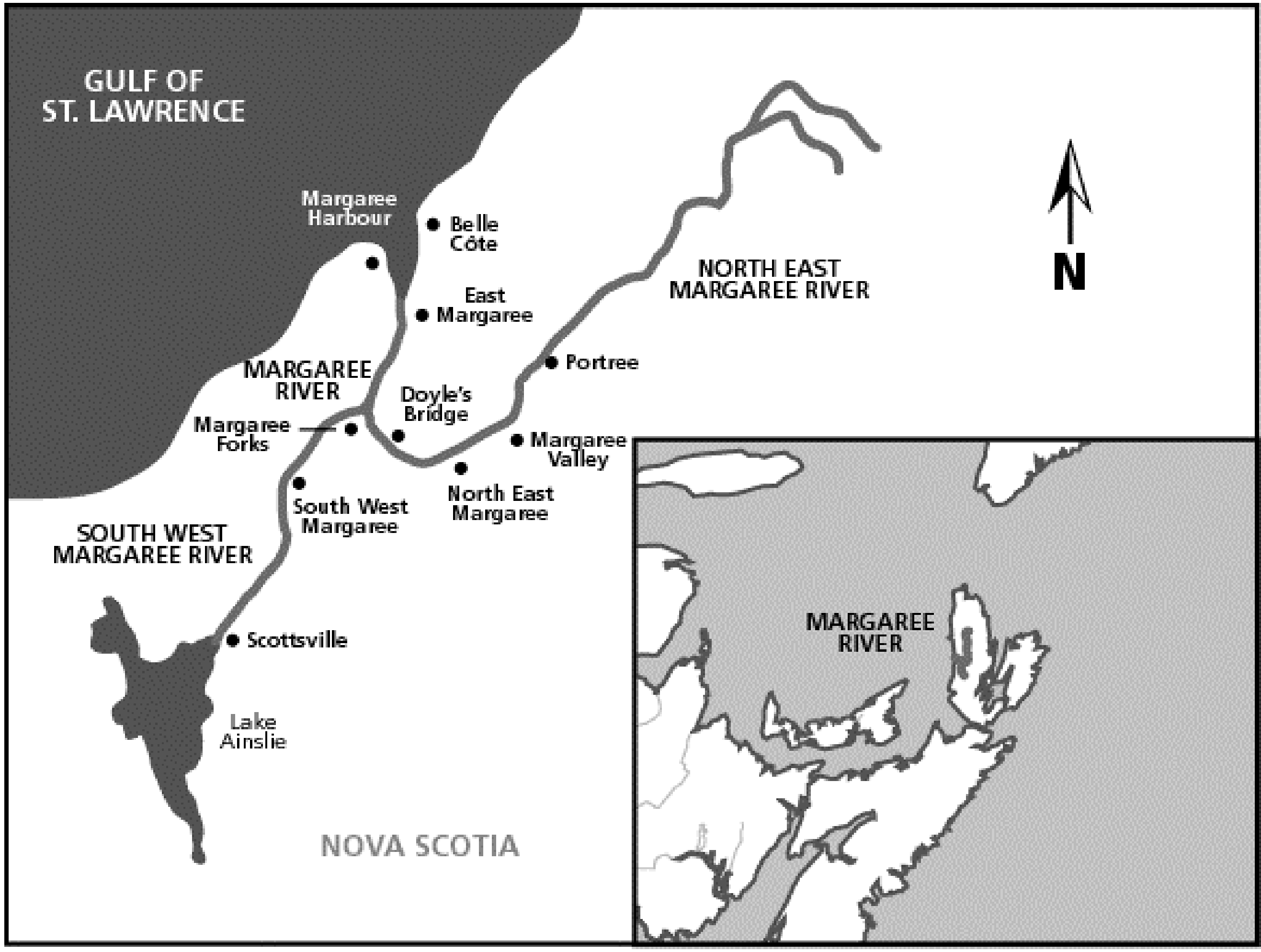
$$\boxed{\frac{1}{\sigma^2} \middle| \theta \sim gamma(a', b')}$$

$$\begin{cases} a' = a + \dfrac{n}{2} \\ b' = b + \dfrac{(Y - X\theta)'(Y - X\theta)}{2} \end{cases}$$

Stock Recruitment issues

- Fishing should not endanger population conservation
- How many spawners are necessary to guarantee salmon population conservation ?
- How does the number of spawners influence on the recruitment ?
- Models of stock/recruitment relationship wanted

Margaree River

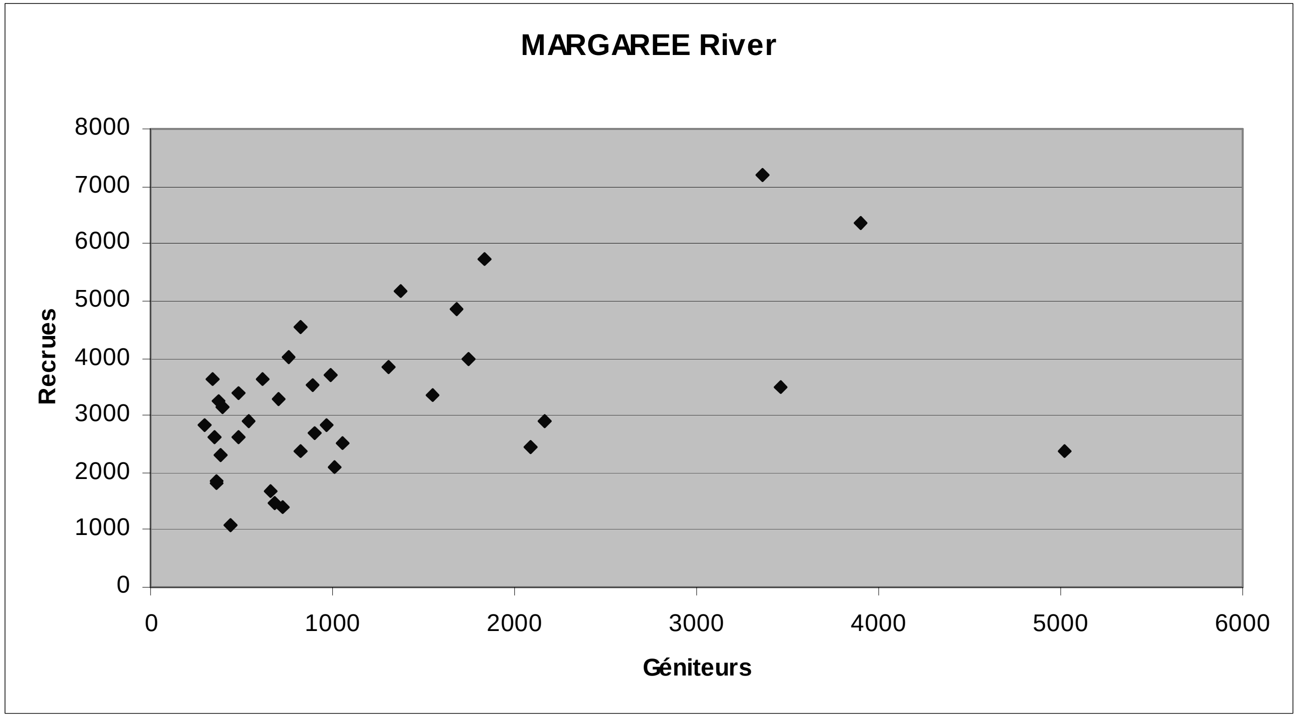


Margaree River

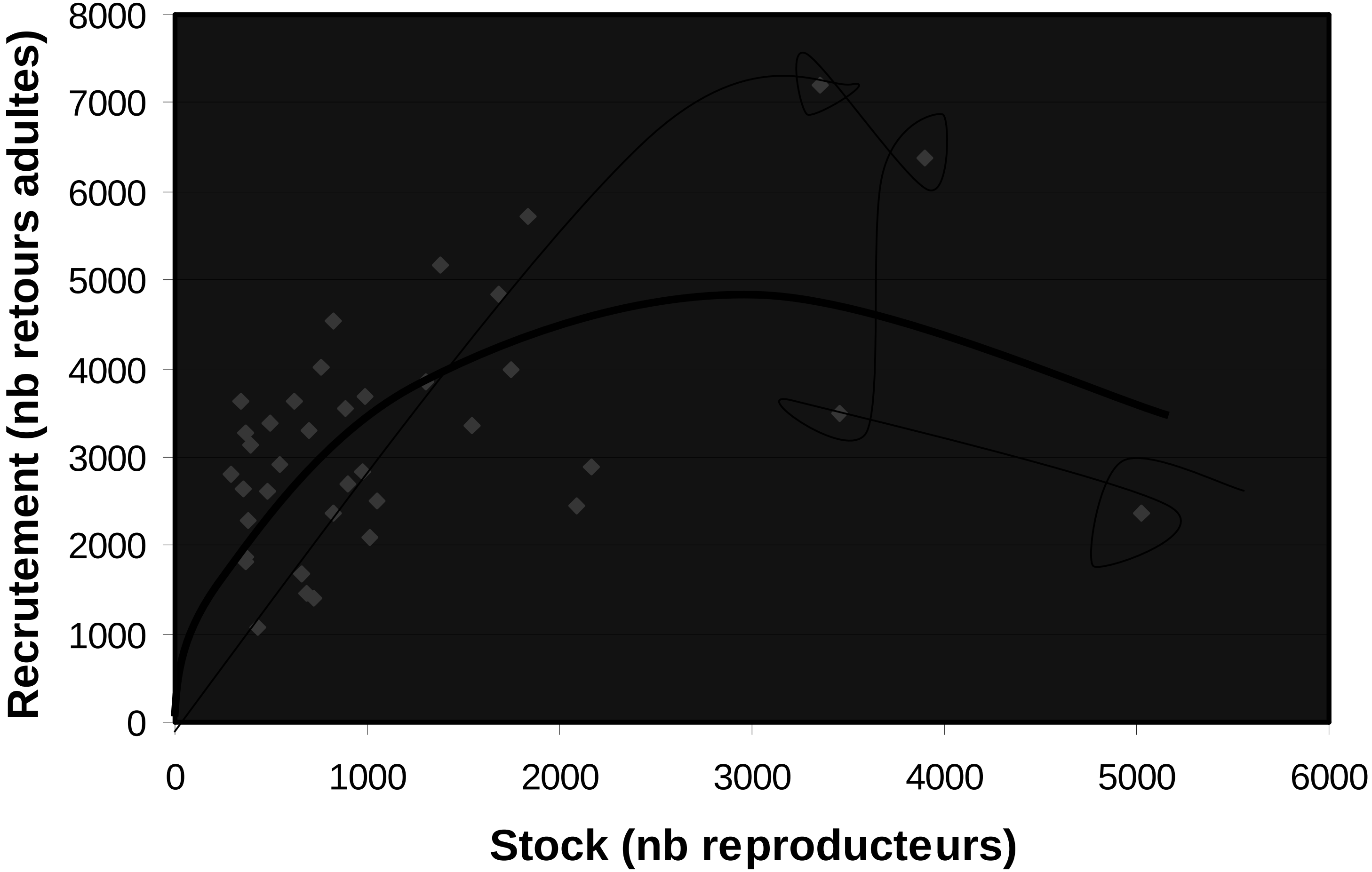
SALMON: The Salmon fishing season begins June 1 and ends October 31 (fly only).Resident fees are set at $36.06. Non-resident anglers must pay $133.19 for a seasonal license or $54.26 for a 7 day license. Bag limit: 2 per day and 8 per season
(only grilse up to 63 cm 24.8 inches)

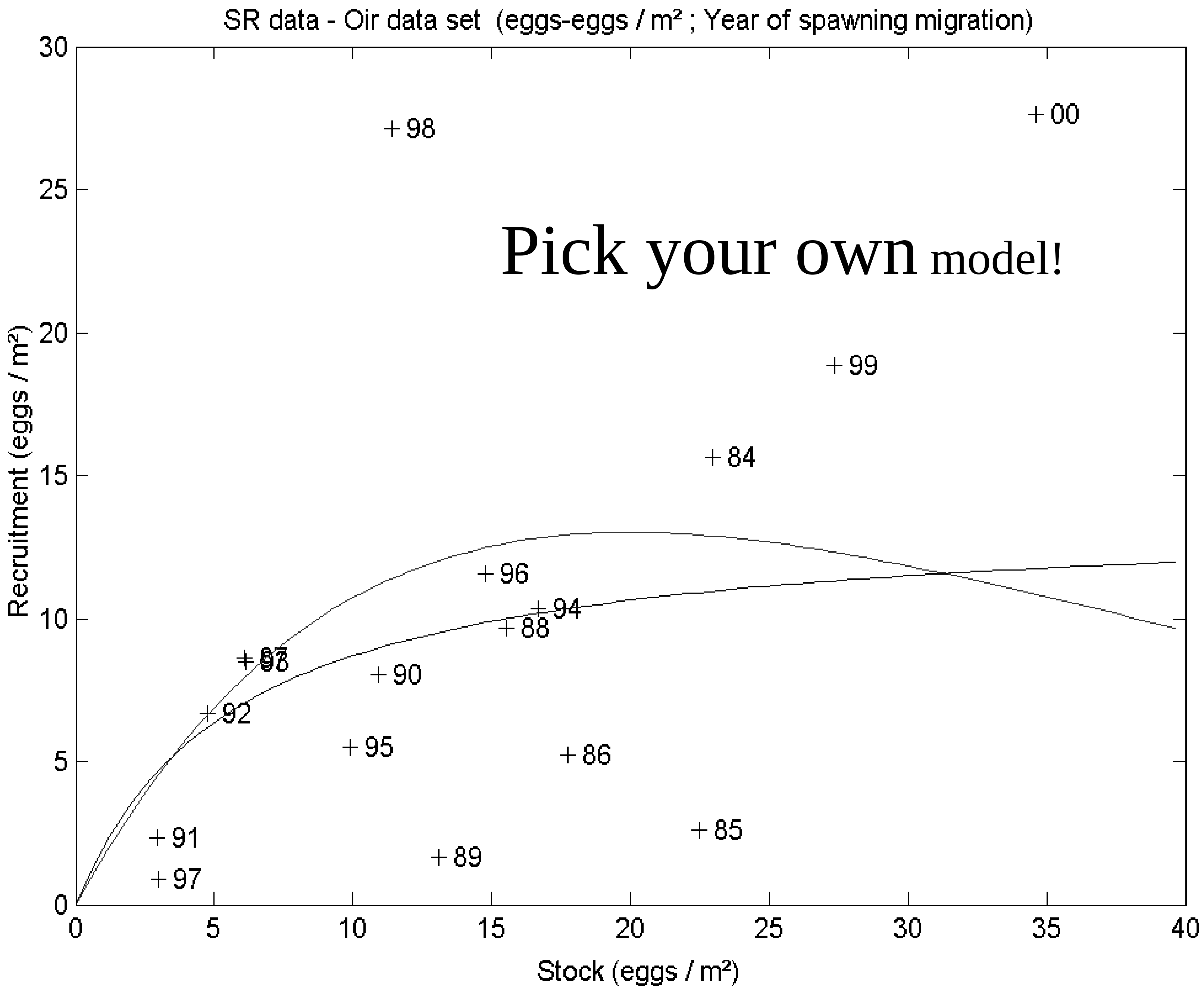
Cohorte	Géniteurs	Recrues
1947	1685	4852
1948	3358	7204
1949	1839	5716
1950	1744	4000
1951	2093	2440
1952	969	2833
1956	486	2616
1957	822	4534
1961	344	3620
1962	1306	3850
1963	887	3538
1964	1053	2515
1965	993	3694
1966	727	1393
1967	1009	2083
1968	828	2378
1969	488	3394
1970	901	2702
1971	351	2630
1972	373	3261
1973	393	3131
1974	436	1066
1975	293	2813
1976	366	1819
1977	538	2909
1978	699	3292
1979	363	1868
1980	681	1462
1981	618	3616
1982	760	4015
1983	657	1688
1984	381	2289
1985	1378	5156
1986	3461	3484
1987	3899	6375
1988	1545	3358
1989	2164	2900
1990	5022	2365

Margaree River



A look at the data of SR Margaree river




SR data - Oir data set (eggs-eggs / m² ; Year of spawning migration)
Pick your own model!
Recruitment (eggs / m²)
Stock (eggs / m²)
0
5
10
15
20
25
30
35
40
98
00
99
84
96
94
88
90
92
95
86
85
91
89
97

Searching for a model

- R as a function of S :
  - a smooth idealized phenomenon R= $\mu_R(S)$,
  - when S ↑, R ↑ (at least at the beginning)

  ⇒ R est fonction de S

- Indeed R is uncertain for many reasons (S is the only explanation available)
  ⇒ stochastic model

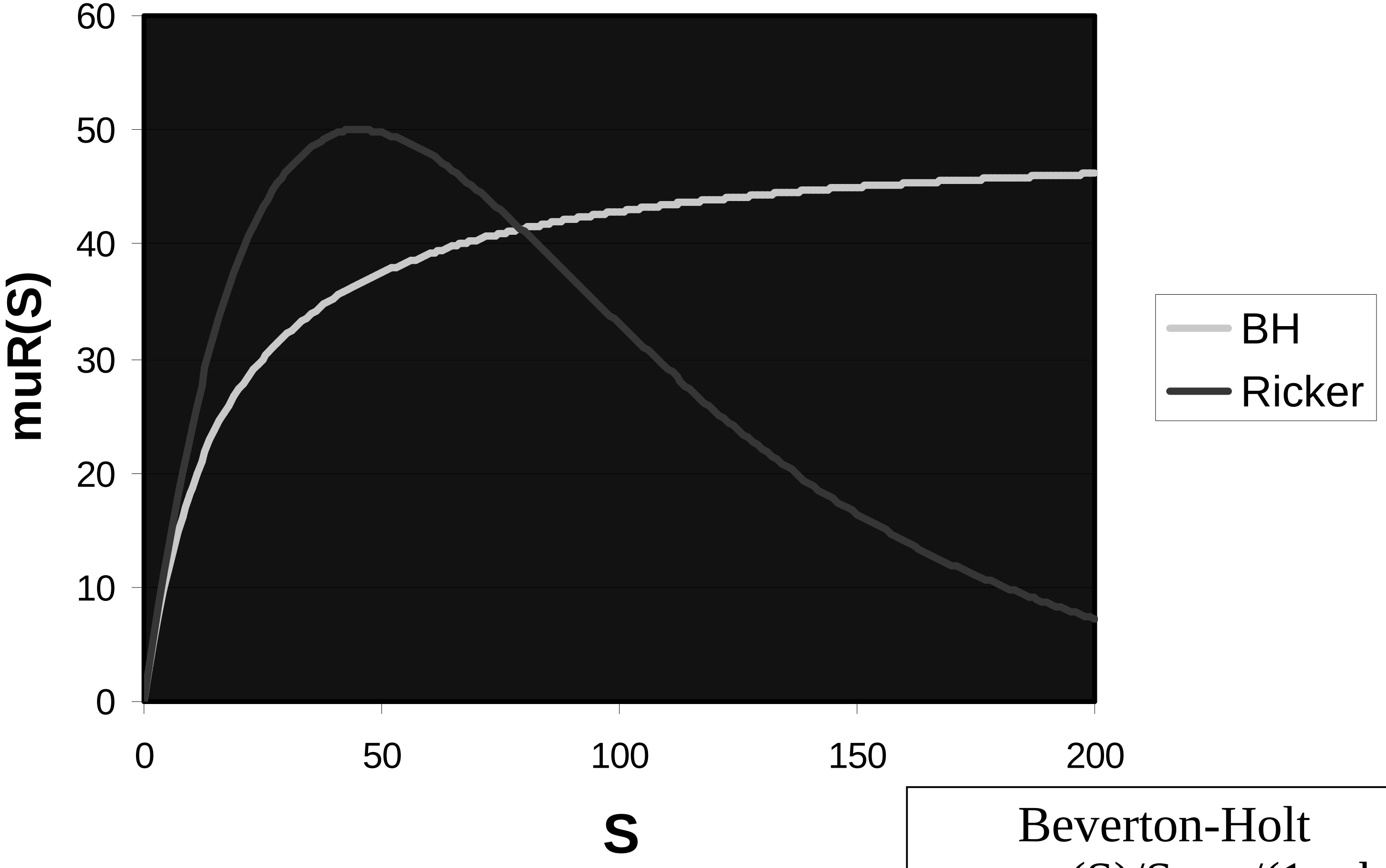
- Math. translation :
  R ~ pdf with caracteristics ($\mu_R(S)$, σ)
  - position parameter $\mu_R(S)$, dispersion parameter σ
  - dispersion parameter σ is not a fonction of S
- $\mu_R(S)$? Which pdf ?

Searching for $\mu_R(S)$

- $\mu_R(S)$?
  - Ecological knowledge : competition for territory
  - R ↑ then S ↑, then R stops or decreases

  ⇒ Feedback control by density
- 2 classical functions :
  - Beverton-Holt
    $\mu_R(S) = aS/(1 + bS)$
    reproduction ratio $\mu_R(S)/S = a/(1 + bS)$
  - Ricker
    $\mu_R(S) = aSe^{-bS}$
    reproduction rate $\mu_R(S)/S = ae^{-bS}$
  - $\boxed{Log(\mu_R(S)/S)=Log(a)-bS}$

Fonctions Ricker et BH



Beverton-Holt
$\mu_R(S)/S = a/(1 + bS)$
Ricker
$\mu_R(S)/S = ae^{-bS}$

Ricker = a hidden linear model

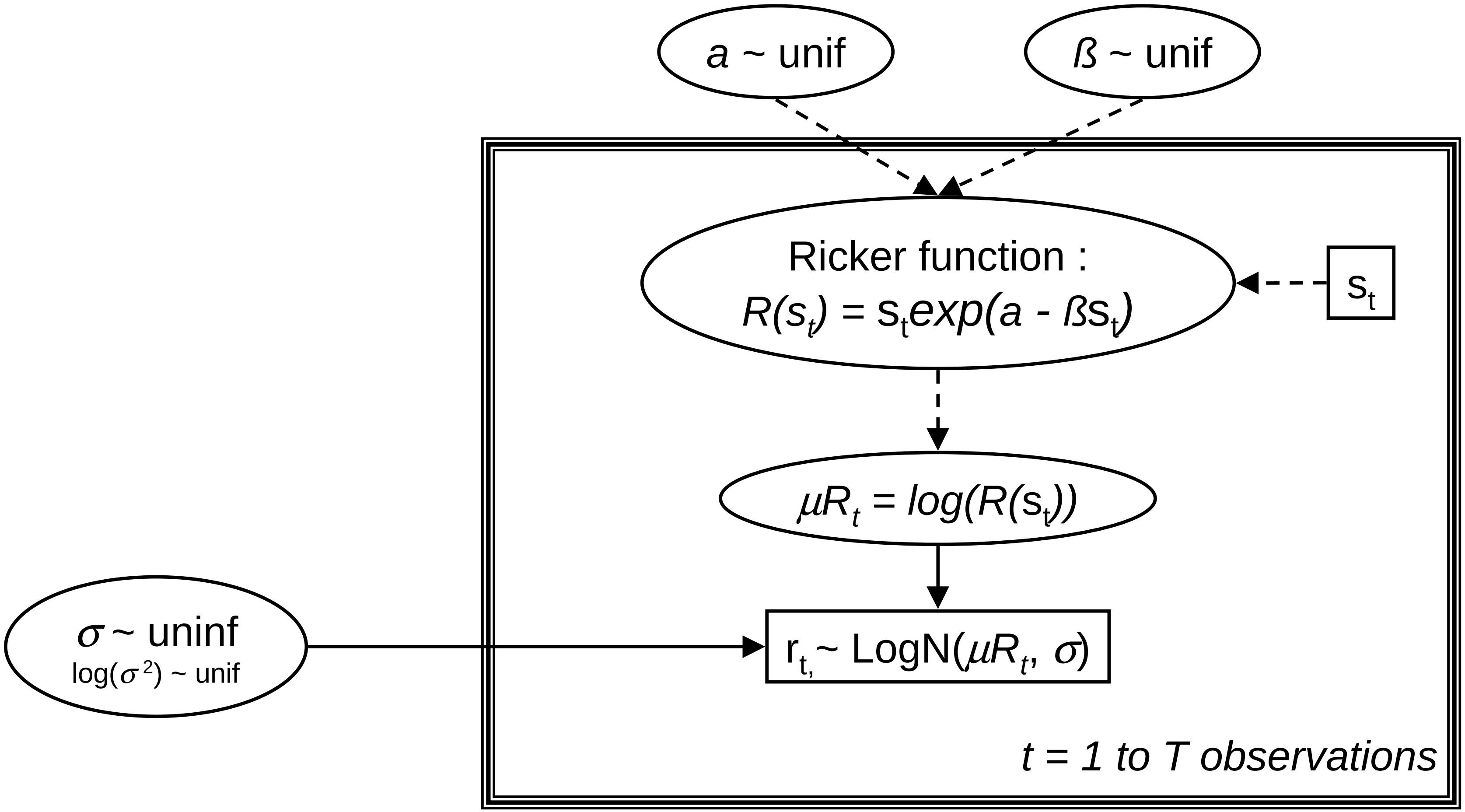
- Summing up
  - $R \sim \log N(\mu_R(S), \sigma)$
  - $\mu_R(S) = aSe^{-bS}$ (Ricker function)
  - deterministic explanatory structure + random effect
- $$R = \mu_R(S)e^{\varepsilon} \quad \varepsilon \sim N(0, \sigma^2) \quad \mu_R(S) = aSe^{-bS}$$

$$\log(R) = \log(a) + \log(S) - bS + \varepsilon$$

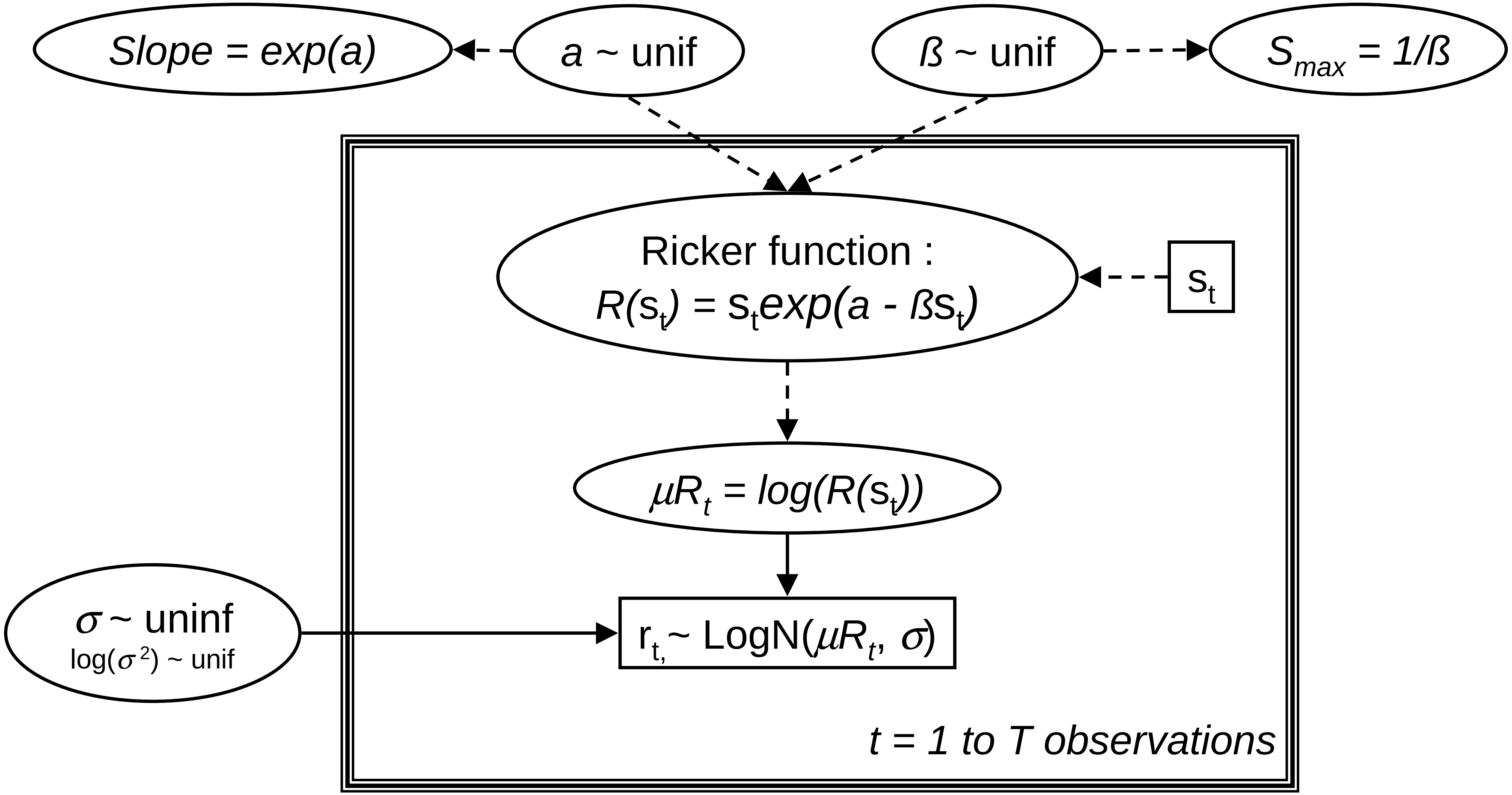
$$y = \log\left(\frac{R}{S}\right) \quad x = S \quad \alpha = \log(a) \quad \beta = -b \quad \theta = (\log(a), b)$$

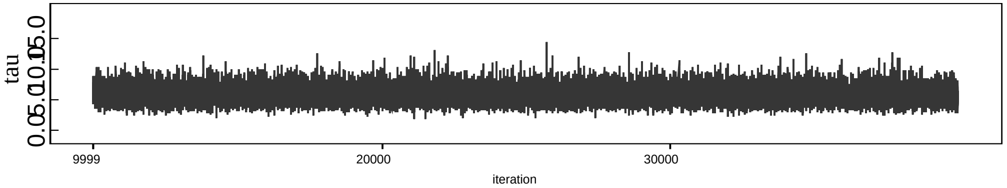
$$y = \alpha + \beta x + \varepsilon \quad \varepsilon \sim N(0, \sigma^2)$$

DAG of the Ricker SR model: classical formulation

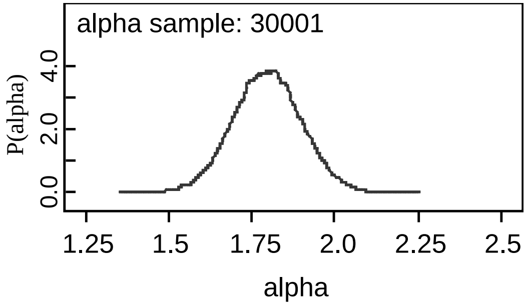


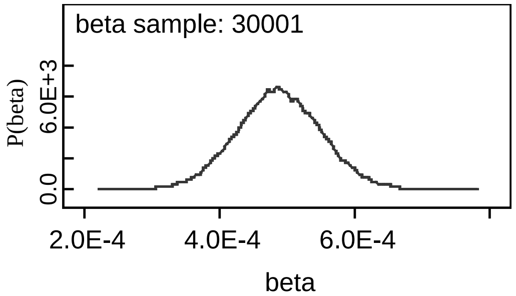
DAG of the Ricker SR model: classical formulation

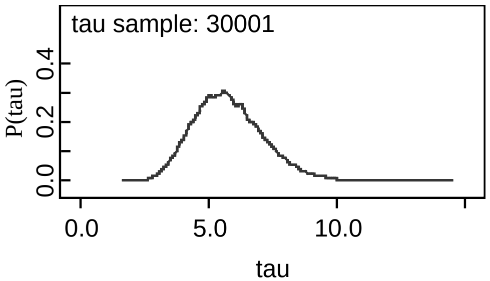


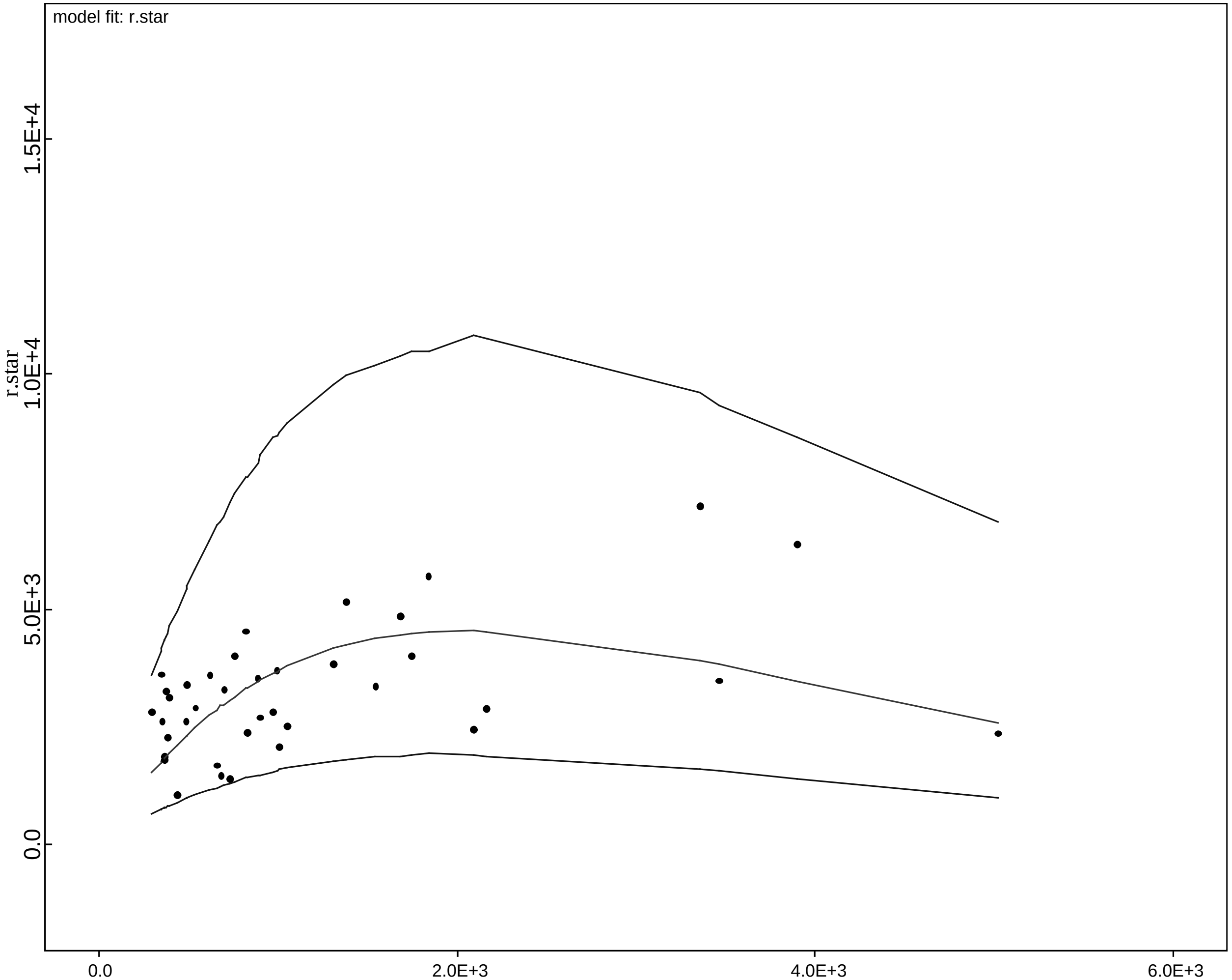


Posterior inference….

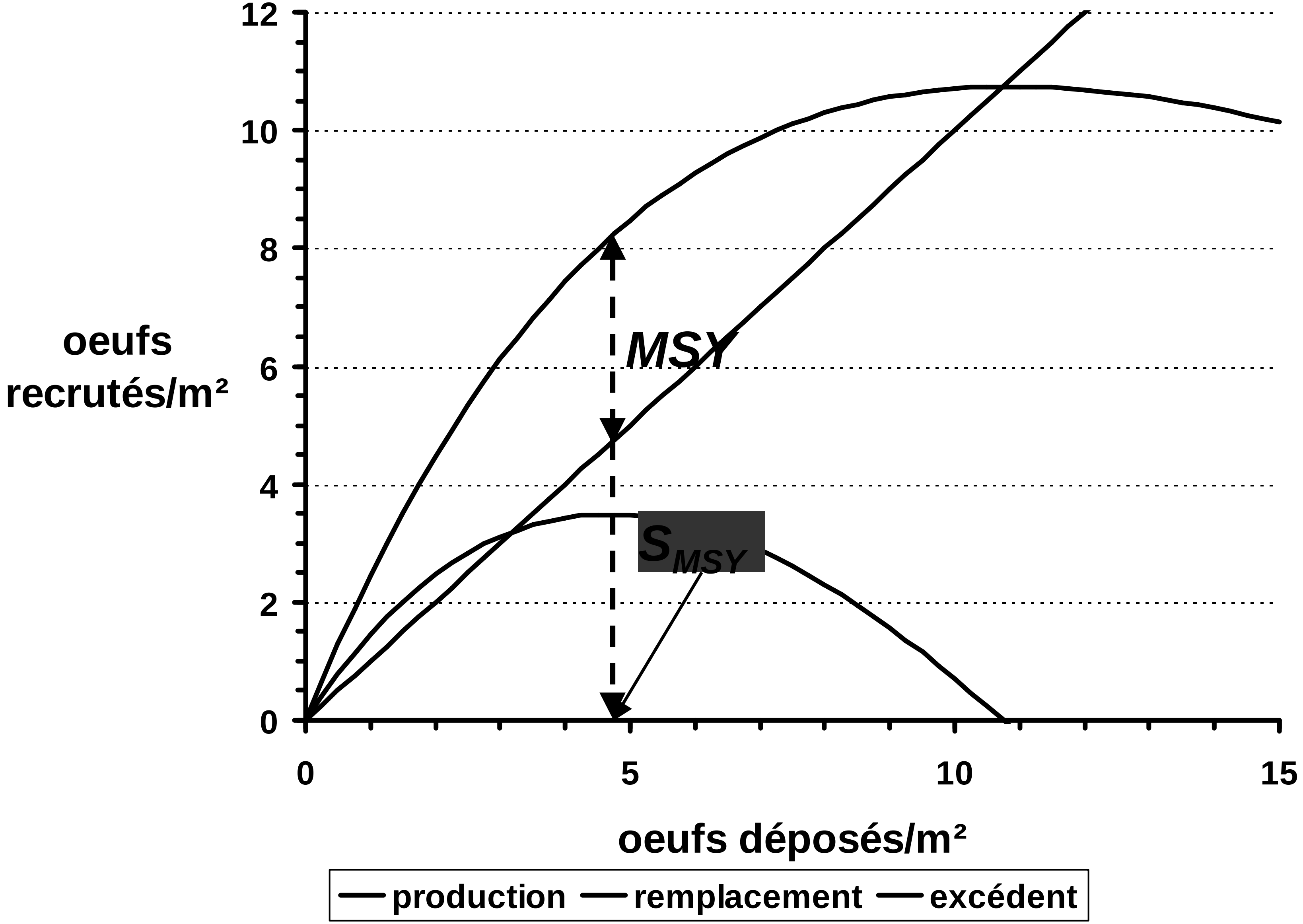






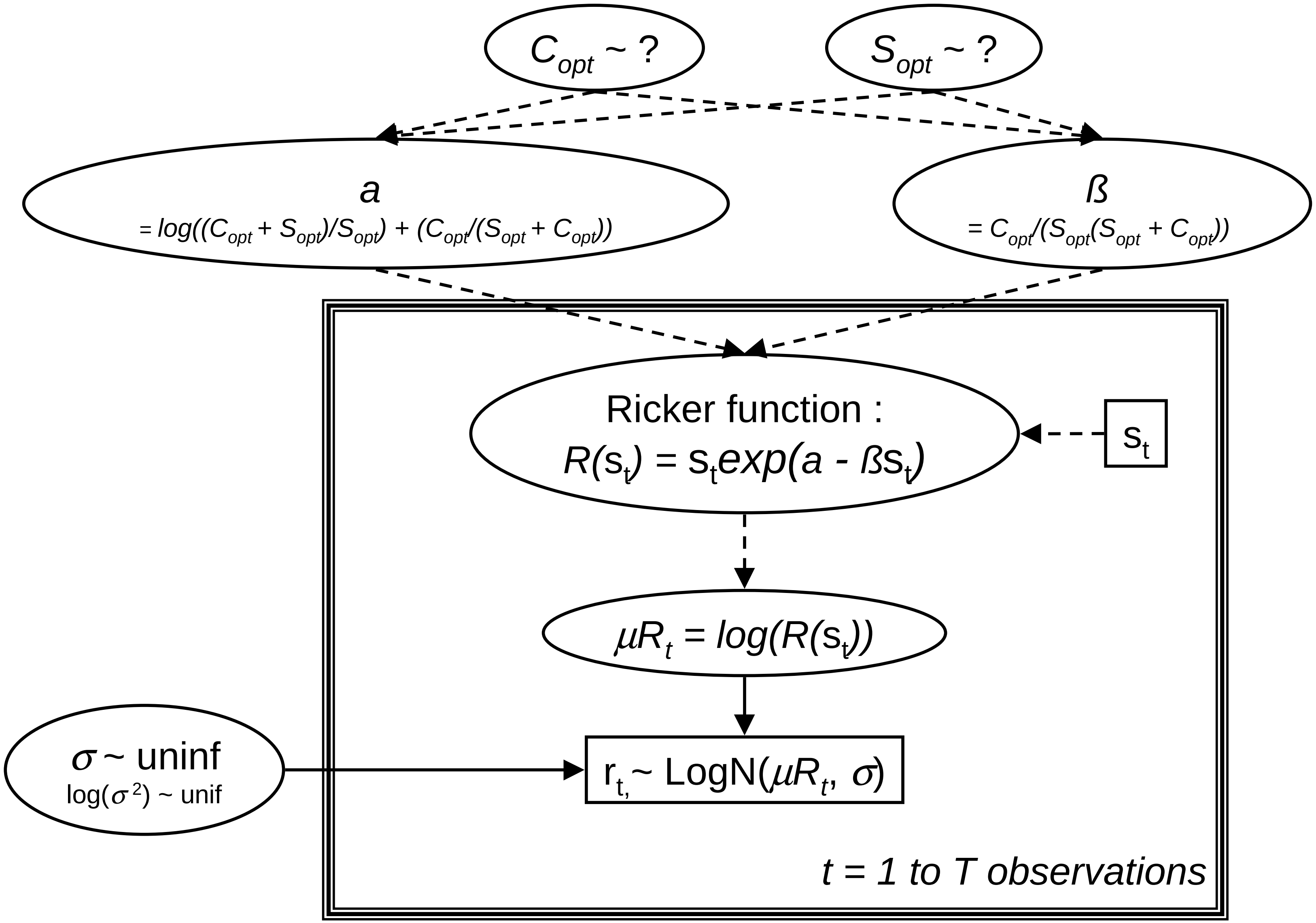


Reference points of the Ricker function



CHANGE of PARAMETERS

Schnute & Kronlund formulation

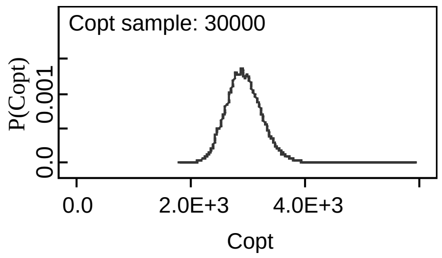


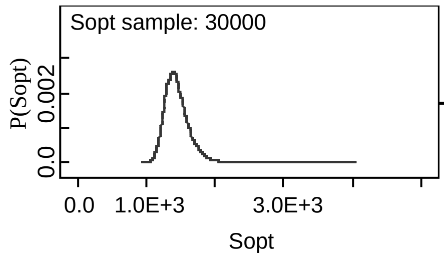
Which priors to assign to C_{opt} and S_{opt}?

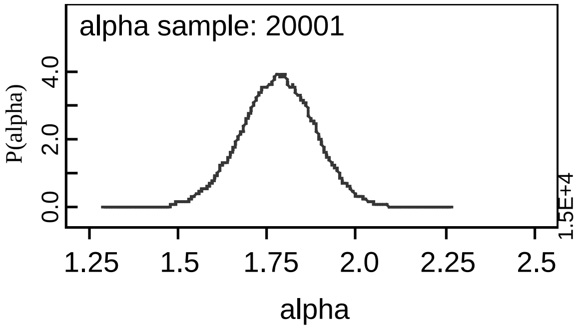
- Option 0 : uniformative priors to α and β
- Option 1: assign a prior equivalent to the reference prior assigned to α and β
- Option 2: propose a little informative prior(s) based on some rationale

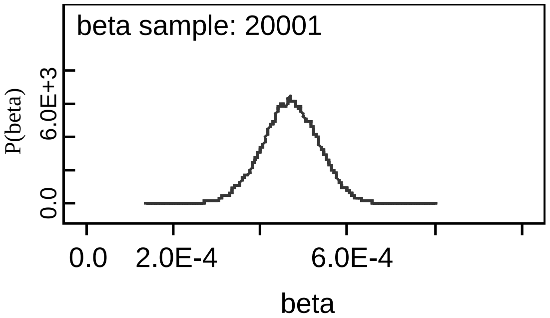
	5%	50%	95%
Copt	2378.0	2926.0	3636.0
Sopt	1170.0	1425.0	1881.0

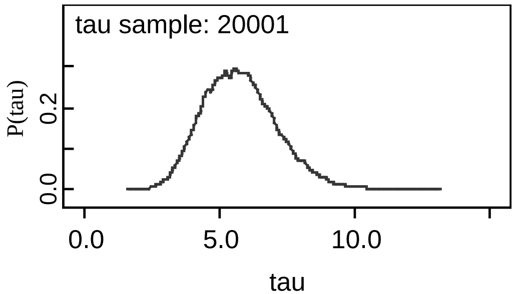
Posterior Analysis 2

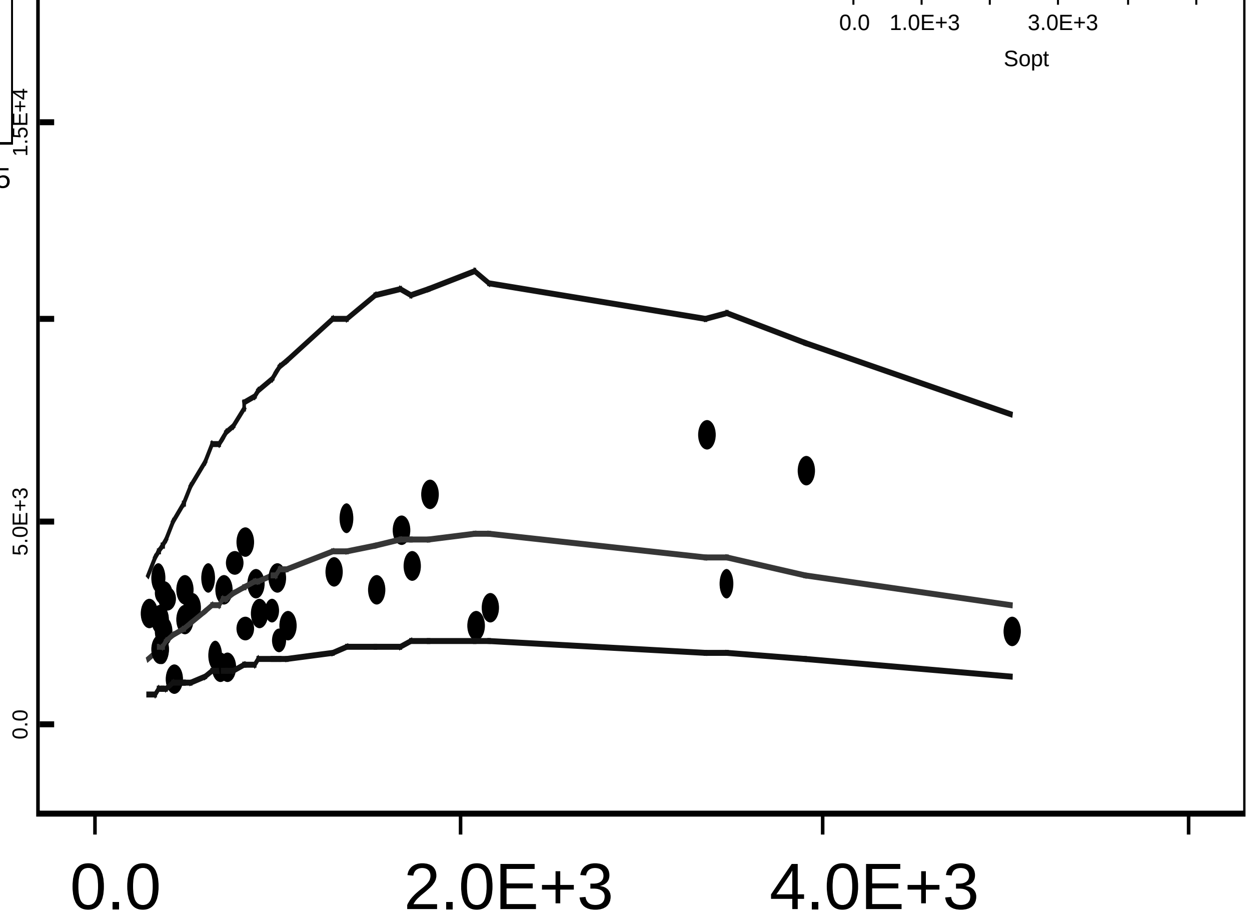












Caution is required when changing parameterization (Rivot et al., 2001): it may carry additional (implicit) hypothesis

- in the case of the Ricker model: the management related parameterization assumes at least a segment of the Ricker curve is above replacement $\Leftrightarrow \alpha > 0 \Leftrightarrow$ slope > 1
- this is a very strong a prior hypothesis for weak population hardly able to replace generations
- Recommendations:
    - always start with classical parameters and priors not excluding $\alpha < 0$
    - assess sensitivity of results to various reasonable priors/parametrization

CHANGE OF ERROR

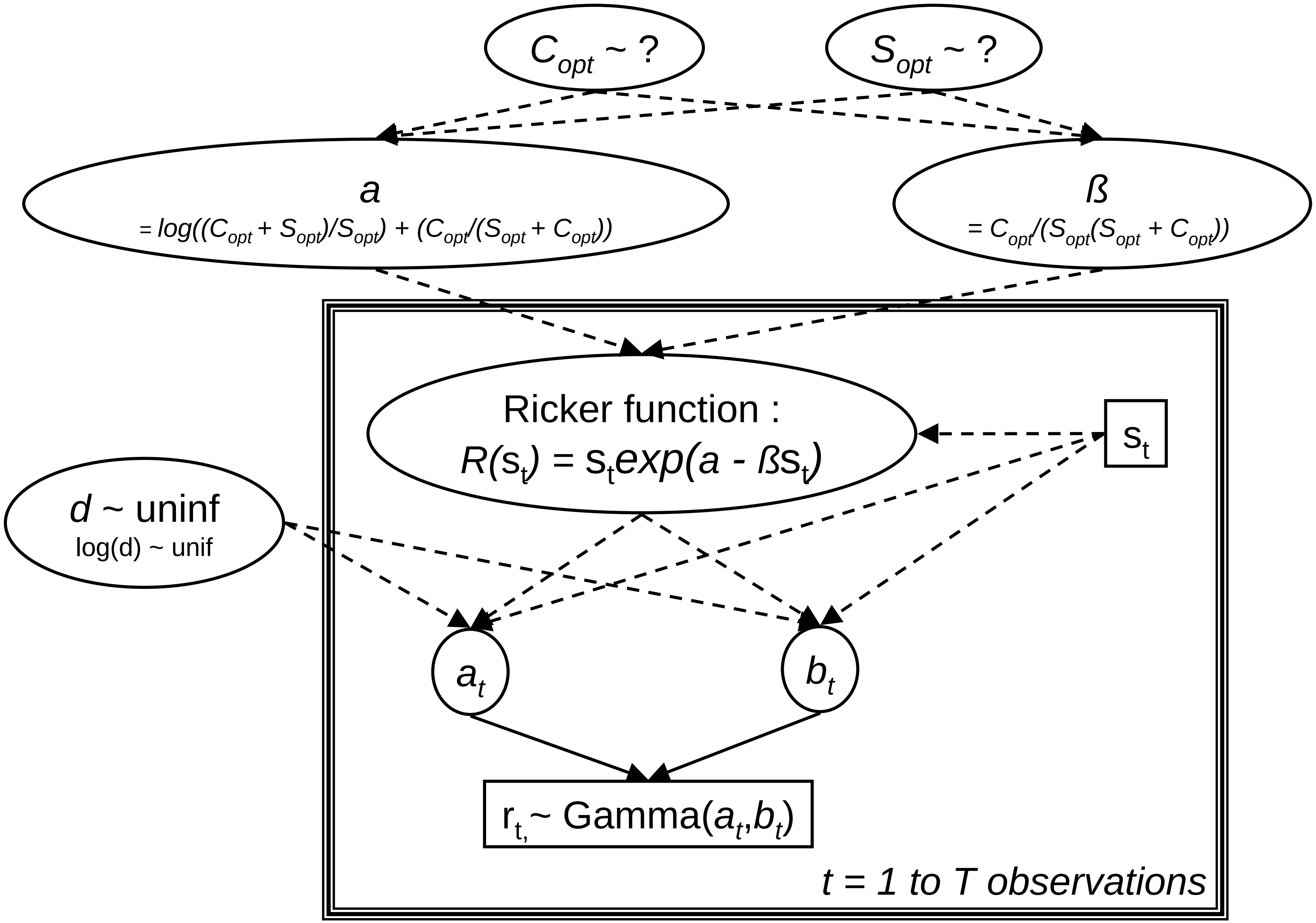
Under the standard lognormal formulation, recruitment variance is proportional to the square of the Ricker function $\Rightarrow$ for $s > S_{max}$, recruitment variance decreases with s

Alternative hypothesis: recruitment variance is proportional to s

Implementation is easy with assuming a gamma distribution for the recruitment (not possible with lognormal)

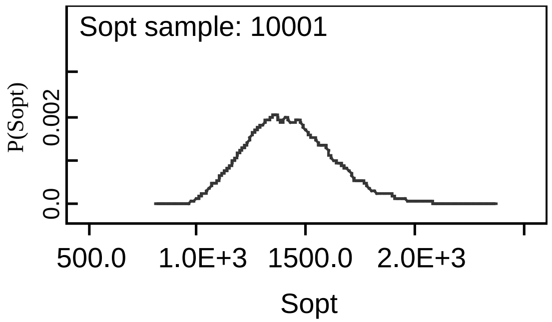
- X ~ gamma(a, b), a = shape, b = inverse scale
  - o $E(X) = a/b$
  - o $Var(X) = a/b^2 = E(X)/b$
- Given the above, propose a formulation for which mean recruitment is the Ricker function and recruitment variance is proportional to the stock

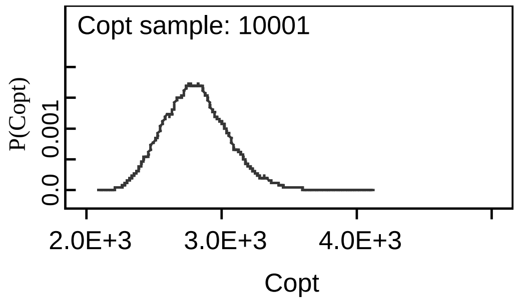
Ricker model DAG : (variance ∝ stock) formulation

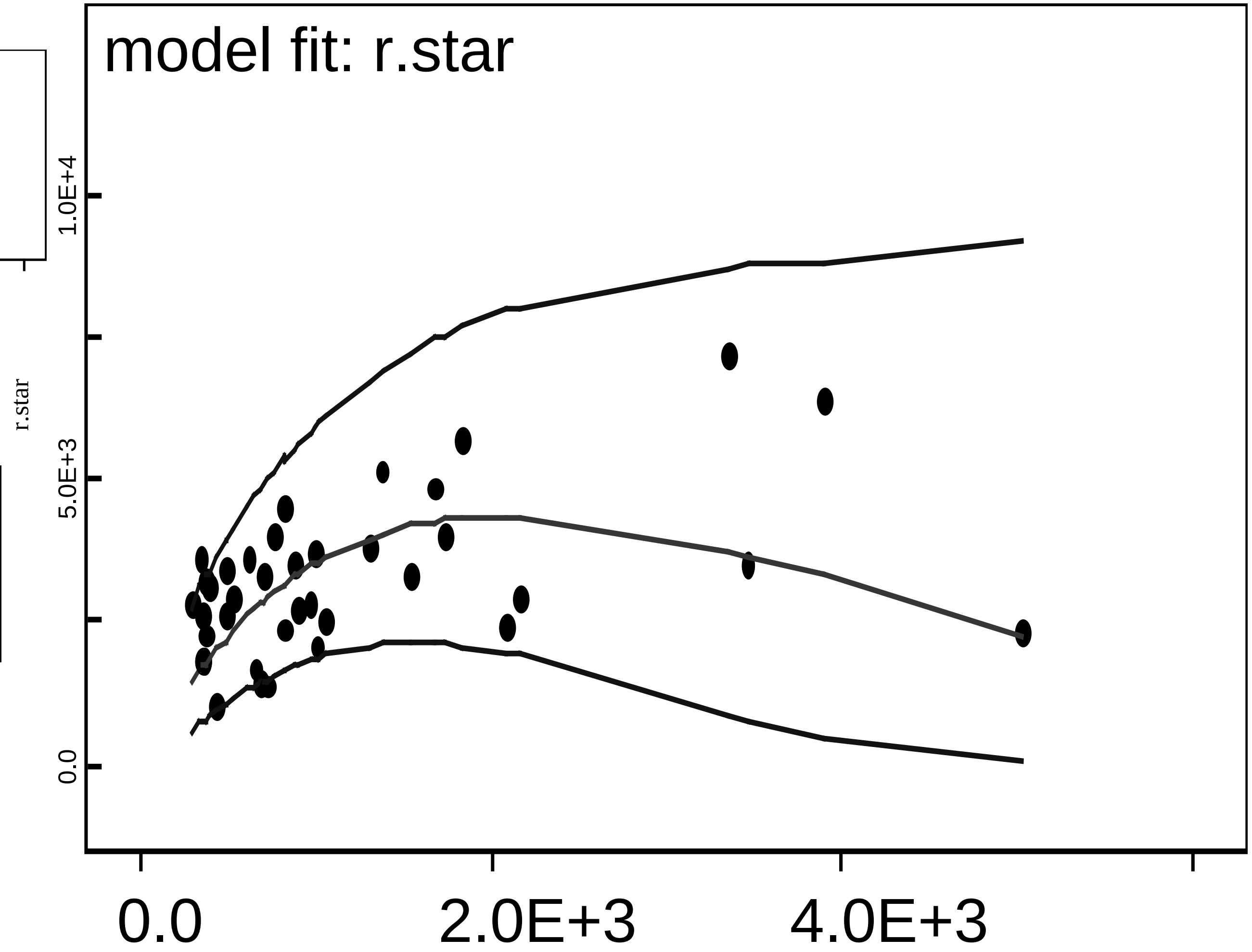


Ricker with gamma errors

	5%	50%	95%
Copt	2379.0	2805.0	3334.0
Sopt	1070.0	1411.0	1870.0







CHANGE OF MODEL Beverton&Holtz =a non linear model

- Summing up
  - R ~ logN(μ_R(S), σ)
  - μ_R(S) = aS/(1 + bS) (Beverton & Holtz type)
  - deterministic explanatory structure + random effect
- $$R = \mu_R(S)e^{\varepsilon} \quad \varepsilon \sim N(0,\sigma^2) \quad \mu_R(S) = \frac{aS}{1+bS}$$

$$\log(R) = \log(a) + \log(S) - \log(1+bS) + \varepsilon$$

$$y = \log\left(\frac{R}{S}\right) \quad x = S \quad \alpha = \log(a) \quad \beta = b \quad \theta = (\log(a), b)$$

$$y = \alpha - \log(1+\beta x) + \varepsilon \quad \varepsilon \sim N(0,\sigma^2)$$

Inference of RS relationships

- Ricker : use the standart linear model theory
  If bayesian, what becomes the prior knowledge (if any) after reparametrization?

- Beverton-Holtz : use non linear techniques

$$[y_1,..y_t,...y_T|\theta,\sigma]=\prod_{t=1}^{T}\frac{1}{\sqrt{2\pi\sigma}}\exp\left(-\frac{1}{2\sigma^2}(y_t-\alpha-\log(1+\beta x))^2\right)$$

If bayesian, which prior [θ,σ]?

Then MCMC…(see WinBugs in a while)

To sum it up...

Modèle			
Ricker	classique	Schnute/Kronlund	
a	6,02558	[redacted]	Sopt
b	0,00049	[redacted]	Copt
SCE	6,17671	6	
Beverton/Holt	classique	Schnute/Kronlund	
a	13,58068	[redacted]	Sopt
b	0,00321	[redacted]	Copt
SCE	5,10412	5	

Summary

- R-S curves : Bayes inference at work Ricker a hidden linear model….
- Reparametrisation : statistical *natural* parameters are seldom *meaningful* for practitionners -> prior modelling
- Role of the error term : what is ε ?
- Discuss the hypotheses
  - Change R/S structure
  - S not known, only observed

Conclusion

Our inferences are not sensitive to priors: the data set is long and informative (good contrast in the spawning stock level)

But

Inferences are more sensitive to modelling hypotheses (recruitment variance proportional to stock or to recruitment)

Which model is best? Model choice by calculating posterior odds of competing models

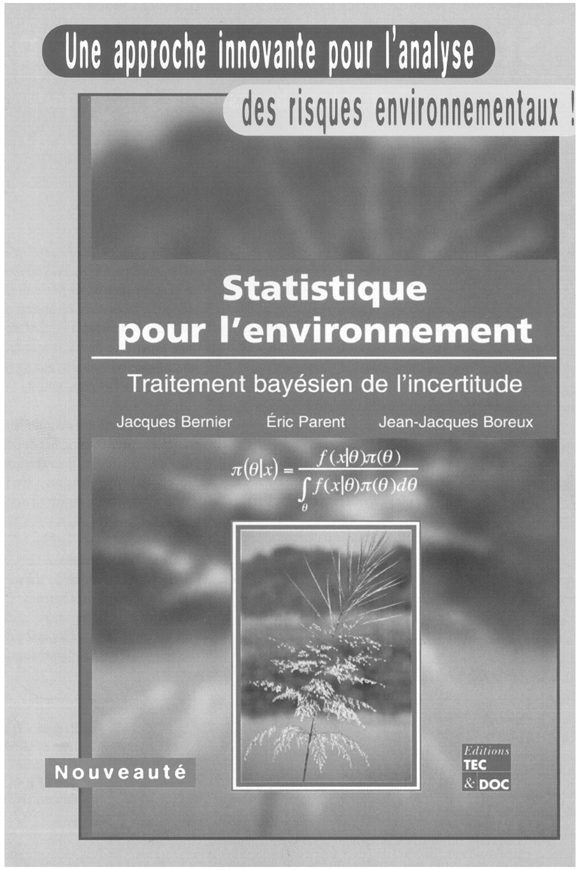
References

Millar R.B., 2002. Reference priors for Bayesian fisheries models. Can. J. Fish. Aquat. Sci., 59: 1492-1502

Rivot E., Prévost E., Parent E., 2001. How robust are Bayesian posterior inferences based on a Ricker model with regards to measurement errors and prior assumptions about parameters? Can. J. Fish. Aquat. Sci., 58: 2284-2297.

Schnute, J. T., and Kronlund, A. R. 1996. A management oriented approach to stock recruitment analysis. Canadian Journal of Fisheries and Aquatic Sciences, 53: 1281-1293

Bernier, J., Parent, E. et J-J Boreux (2000), Statistique pour l’environnement : traitement bayésien de l'incertitude, Eds Tec & Doc (Lavoisier), Paris (363 p.).



- Texte « pédagogique »
- Mathématiques
  - Niveau « bac »
  - + quatre intégrales
- Prix : 65 €