Lecture 10: Bayesian modelling of time series

Outline of lecture 10

- What is Bayesian statistics?
- What is a state-space model? Or why use Bayesian statistics?
- Known vs. unknown distributions: BTS vs. BUGS
- What is simulation?

What is Bayesian statistics?

My favourite definition: "Everything that you think that frequentist statistics is"(!)[1]

⇨ A different way of thinking – appealing. Much more intuitive and straightforward

⇨ Instead of asking: What is the likelihood of this data point given the model (frequentist), the Bayesian ask: What is the likelihood of the model given this data point?

Short history:

o "Normal" (classical, frequentist) statistics formalised in the early 20th century (Karl Pearson, Ronald Fisher et al.), became dominant.

o Bayesian philosophy developed by Reverend Thomas Bayes in late 18th century

o Revival of Bayesian statistics in late 20th century due largely to computational advances (software and computing power)

---

[1] M. Kittilsen, pers com.

Classical/Frequentist vs. Bayesian statistics

Concept	Frequentist	Bayesian
Probability	Long-run relative frequency with which an event occurs in m|any repeated similar trials. Objective quantity	Measure of one's degree of uncertainty about an event. Subjective quantity
Inference	Evaluate the probability of the observed data, or data more extreme, given the hypothesized model (H_0)	Evaluating the probability of a hypothesized model given observed data
95% C. Interval	A 95% *Confidence* Interval will include the fixed parameter 95% of the trials under the null model	A 95% *Credibility* Interval contains the parameter with a probability of 0.95.
"Range"	The P-value (of significance) measures the probability of getting a result at least as extreme as the one observed, given that the null hypothesis is true.	The probability of any given model can be evaluated given observed data. Unobserved data are irrelevant.

Formal framework of Bayesian Statistics

Bayes's theorem (entirely uncontroversial) states that the probability that event A occurs, given that event B has occurred, is equal to the probability that both A and B occur, divided by the probability of the occurrence of B:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Now setting A as a parameter, a collection of parameters, *i.e.*, the hypothesis (θ) and B as the obtained data (x):

$$P(\theta|x) = \frac{P(\theta \cap x)}{P(x)} = \frac{P(x|\theta) \times P(\theta)}{P(x)}$$

$P(\theta|x)$ is the *Posterior* (probability) of obtaining a parameter estimate θ, given the data obtained.

$P(x|\theta)$ is the *likelihood* of obtaining the data under the hypothesis (the same quantity as in frequentist statistics)

$P(\theta)$ is the *Prior* probability of θ

$P(x)$ is the probability of obtaining the data under all admissible parameter estimates (essentially a scaling constant)

"Posterior = prior x likelihood"

Mighty Joe and Herman
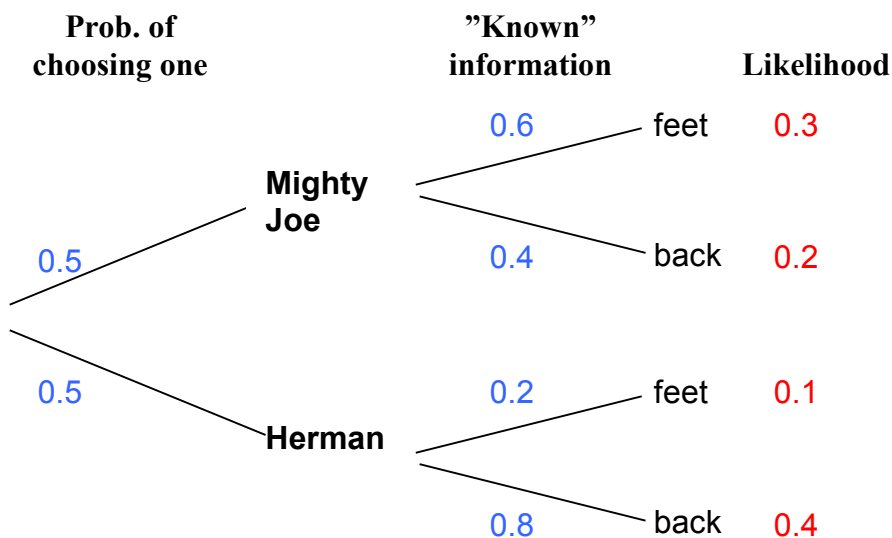
– An example

Two paper frogs:

Mighty Joe:

Lands on feet 60%

Herman:

Lands on feet 20% of times

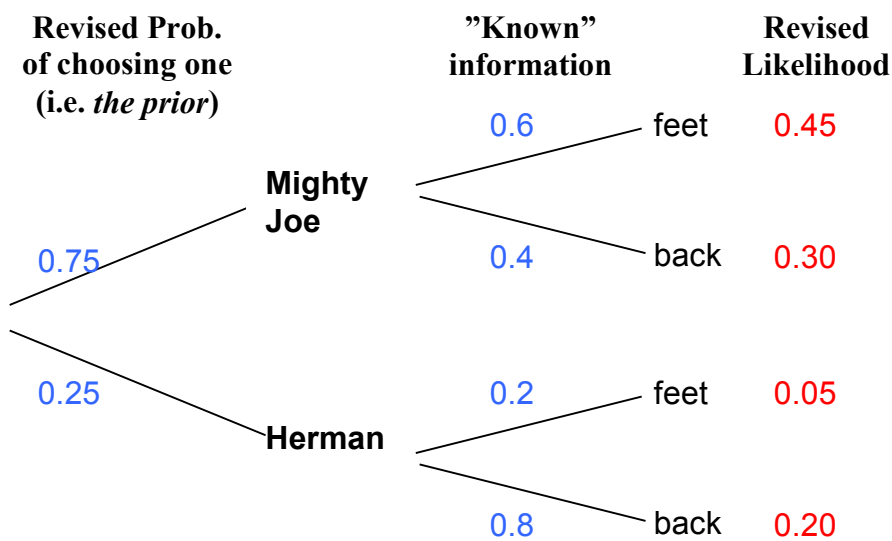- We pick one frog, "jump" it. It lands on its feet. What is the probability it is Mighty Joe?

Prob. of choosing one		"Known" information		Likelihood
		0.6 — feet		0.3
	Mighty Joe			
0.5		0.4 — back		0.2
0.5		0.2 — feet		0.1
	Herman			
		0.8 — back		0.4

We frame this in the Bayesian setting:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$, and in our example:

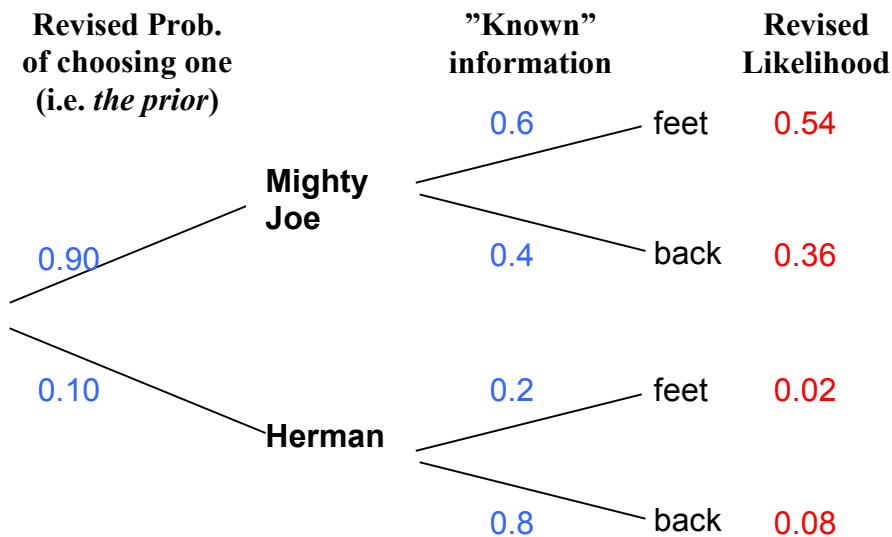$$P(Joe|feet) = \frac{P(Joe \cap feet)}{P(feet)} = \frac{0.3}{0.3 + 0.1} = 0.75$$

Thus, the probability that this was Mighty Joe is 75%.

- We "jump" the frog again. Again it lands on its feet. How can we *Update* our belief whether or not this is Mighty Joe?

Revised Prob. of choosing one (i.e. *the prior*)		"Known" information		Revised Likelihood
	Mighty Joe	0.6	feet	0.45
0.75		0.4	back	0.30
0.25	**Herman**	0.2	feet	0.05
		0.8	back	0.20

$$P(Joe|feet) = \frac{P(Joe \cap feet)}{P(feet)} = \frac{0.45}{0.45 + 0.05} = 0.90$$

▪ We "jump" the frog a third time. This time it lands on its back. How does that influence our belief in this being Mighty Joe?

Revised Prob. of choosing one (i.e. *the prior*)		"Known" information		Revised Likelihood
	Mighty Joe	0.6	feet	0.54
0.90		0.4	back	0.36
0.10	**Herman**	0.2	feet	0.02
		0.8	back	0.08

$$P(Joe|feet) = \frac{P(Joe \cap feet)}{P(feet)} = \frac{0.36}{0.36 + 0.08} \approx 0.82$$

We have achieved:

- A probabilistic interpretation of the problem at hand: "It is 82% change that the chosen frog is Mighty Joe".

- Demonstrated general approach to Bayesian statistic

General approach in Bayesian statistics:

- Use available information to develop a prior.
- Get new data
- Find posterior
- Update prior
- Get new data

This is an appealing framework of statistics.

Is this why we should become Bayesian?

Not really – My reason is pragmatic, and it involves a short detour to state-space models.

State-space models

(Chatfield C. 1996. *The analysis of time series: an introduction.* Chapman & Hall, London: Chapter 10)

Scientists (and even more so biologists) discover that measurement of any signal is contaminated by noise:

Observation = signal + noise

It has been found appropriate to decompose these processes into two: **state variables** (state vector) and **observation variables** (observation vector). One (of many notations) is:

Observation equation:

$$Y_t = G_t X_t + W_t$$

Transition equation (process equation, state equation):

$$X_t = F_t X_{t-1} + V_t$$

Y_t: Observations

X_t: The (hidden) variable of interest, the biological data (*e.g.*, the population size etc.)

W_t: The noise of the observation process

G_t: The transition matrix of the observation process, the collection of parameters

F_t: The transition matrix of the deterministic process, the collection of parameters

V_t: The noise of the population process

An example ("10.1.1. The steady model"): The linear regression

Observation equation:

$$Y_t = X_t + \varepsilon_t$$

Transition equation (process equation, state equation):

$$X_t = X_{t-1} + \beta + \upsilon_t$$

Y_t is our observations.

X_t is our is the level

β is the increment (the trend).

ε_t and υ_t is independent mutually independent noise.

State-space models have been known for a long time, and they are intuitively attractive.

They have appeared towards the back of (time series) text books, software and methods for applications have been missing.

Estimation of state-space models has been by way of the **Kalman Filter**. A Kalman Filter is a recursive set of equations to update the estimated parameters.

We see that the idea of updating in the Kalman Filter is related to the Bayesian approach described earlier. Indeed, the theory behind the Kalman Filter is Bayesian.

Until recently, it has been very challenging to actually perform the Kalman Filter procedure.

Now, increased computational power and software (i.e., BUGS) have made it possible for a wide variety of scientist to estimate state-space models.

Known vs. Unknown distribution (=> Likelihood)

If we can assume that the likelihood is known, i.e., that it can be written out, the recursive equations can be explicitly calculated.

Example: Airpass.dat

The number of air passengers can be expressed as a growth model (the upper equation is the observation model):

$$Y_t = \mu_t + \nu_t$$

Observation equation

$$\mu_t = \mu_{t-1} + \beta_{t-1} + \omega_{t1}$$
$$\beta_t = \beta_{t-1} + \omega_{t2}$$

State equations

where Y_t is the observed number of passengers,

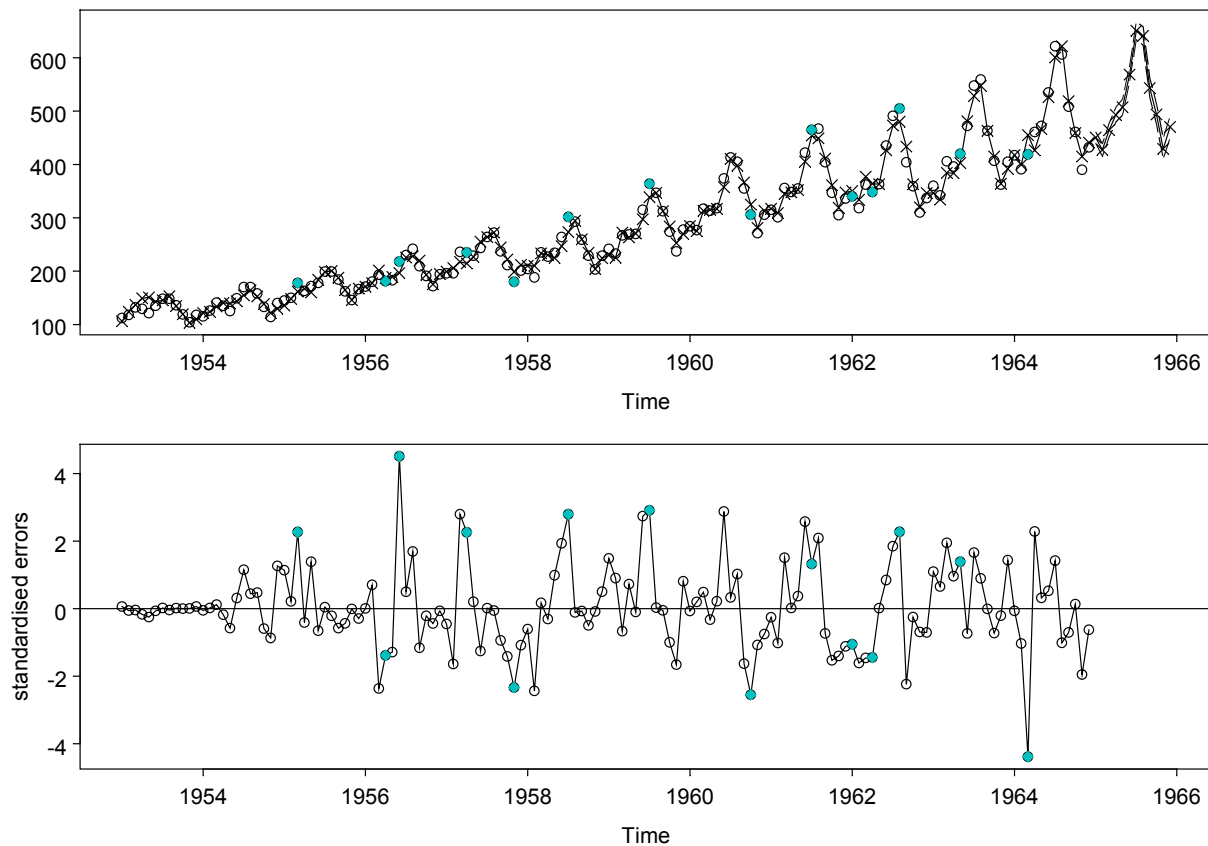μ_t is the level

β_t is the growth/increment in passenger level (allowed to vary).

ν_t, ω_{t1} and ω_{t2} is independent mutually independent noise.

In addition the seasonal component is removed.

These equations have been fitted by the splus codes "`bts`" based on West M. & Harrison J. 1997. *Bayesian forecasting and dynamic models*. Springer, New York.
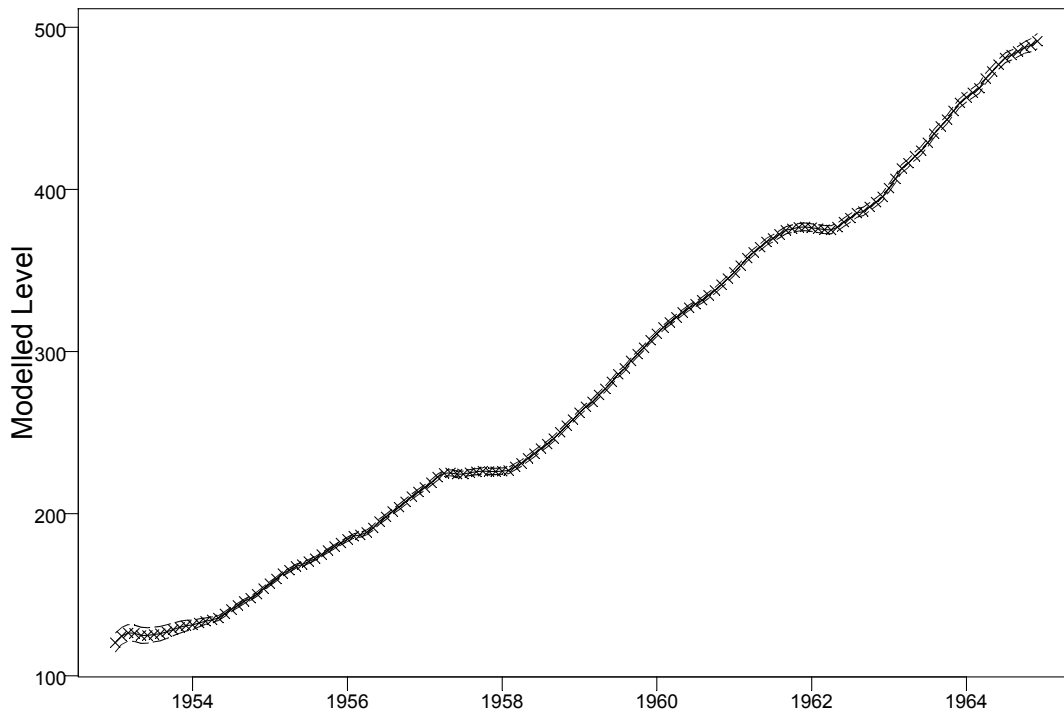
Forecast and residuals

Upper is forecast. Lower is residuals (the blue dots is "interventions", which is another feature of applied Bayesian forecasting philosophy not treated here).
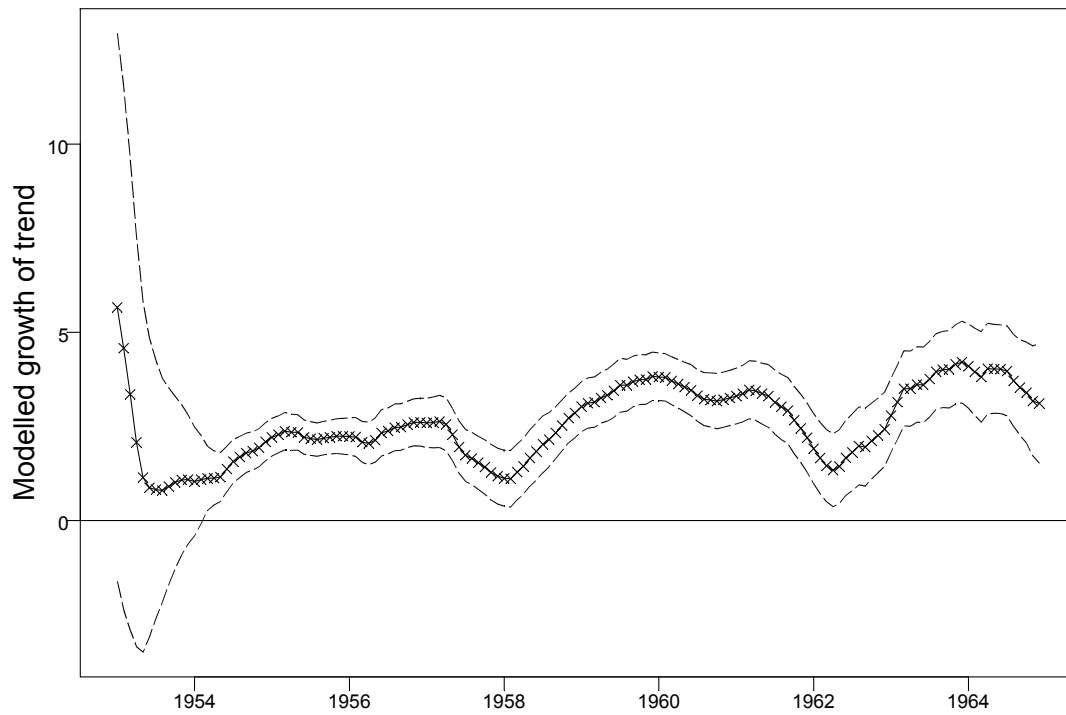
-> Very nice fit. As growth is allowed to vary, a very good fit can be achieved.
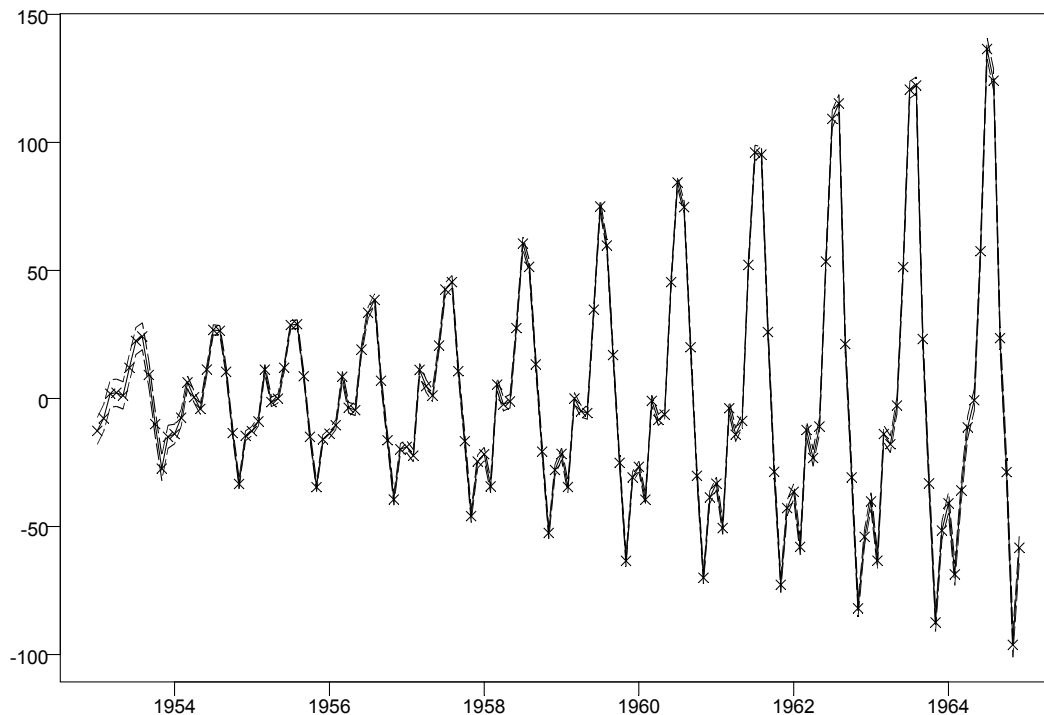
Level:

-> Almost a linear trend (but allows for stagnant periods).

Growth/increment

-> What creates the level.

Season

-> Here the variance is taken care of….

However, when we cannot write down the likelihood, we need to simulate. This is were the revolution has taken place: (Win)BUGS!

BUGS and time series modelling

(Bayesian inference Using Gibbs Sampling)

Freely available: http://www.mrc-bsu.cam.ac.uk/bugs/welcome.shtml

- A program package for simulation of data, using certain schemes ("Gibbs" and "Metropolis-Hastings") to sample from distributions. Using the simulated data to draw inference of the parameters, given the data

- Taking advantage of the Markov Chain property (*i.e.*, a set up in which all information up to time t is contained in the information for time t-1).

- Using the Monte Carlo principle[2] to obtain non-random information ("Integration")

- If the distribution is known, BUGS is not necessary (the problem could also be solved by, *e.g.*, maximum likelihood methods). However, the results will be overall in agreement.

- If the distribution is not correctly specified using ml-methods, simulation may lead to a different (an probably more correct) result.

- Complex likelihoods can be specified

---

[2] The method is called after the city in the Monaco principality, because of roulette, a simple random number generator. The term 'Monte Carlo' was introduced by von Neumann and Ulam during World War II, as a code word for the secret work at Los Alamos (M. Kittilsen, pers.com)
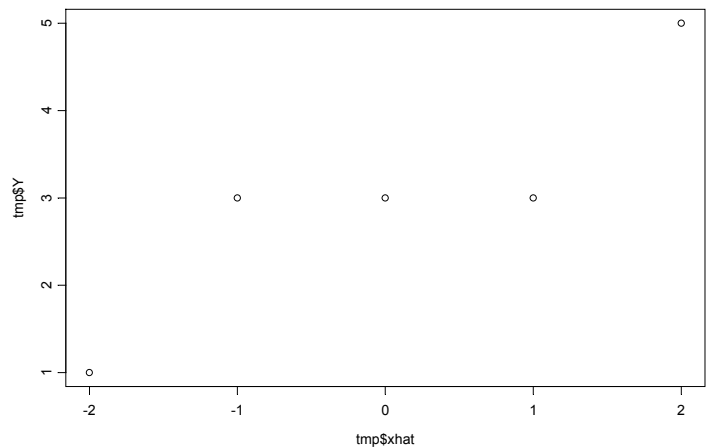
WinBUGS example I: Linear regression
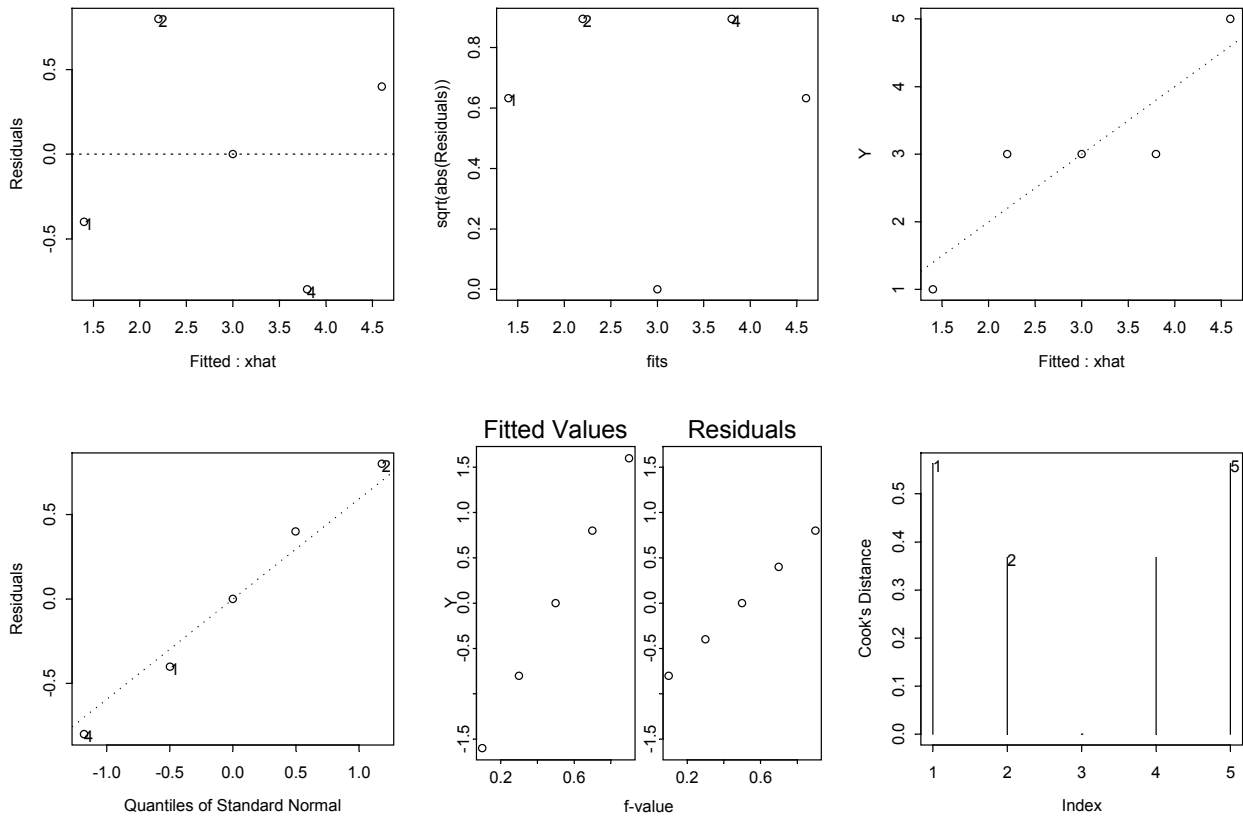
1.   Traditional frequentist analysis

```
tmp <- list(x = c(1, 2, 3, 4, 5), Y = c(1, 3, 3, 3, 5), N = 5)
tmp$xhat <- tmp$x-mean(tmp$x)
tmp
$x:
[1] 1 2 3 4 5
$Y:
[1] 1 3 3 3 5
$N:
[1] 5
$xhat:
[1] -2 -1  0  1  2

plot(tmp$xhat, tmp$Y)

tmp.lm <- lm(Y ~ xhat,
data=tmp)
```

```
par(mfrow=c(2,3))
plot(tmp.lm)
```



```
summary(tmp.lm)

Coefficients:
            Value Std. Error t value Pr(>|t|)
(Intercept) 3.0000 0.3266    9.1856  0.0027
      xhat 0.8000 0.2309     3.4641  0.0405

Residual standard error: 0.7303 on 3 degrees of freedom
Multiple R-Squared: 0.8
F-statistic: 12 on 1 and 3 degrees of freedom, the p-value is
0.04052
```

We estimate the mean to be 3 (interval: 3±1.96*0.3266: 2.36 to 3.64) with the slope being 0.8 (sd: 0.23).

2. Using WinBUGS (from the User manual):

WinBUGS13

File   Tools   Edit   Attributes   Info   Model   Inference   Options   Doodle   Window   Help

i.) A model is specified:

$$Y_i \sim \text{Normal}(\mu_i, \tau) \qquad\qquad \text{\# Observation equation}$$
$$\mu_i = \alpha + \beta(x_i - x.bar) \qquad\qquad \text{\# System equation}$$
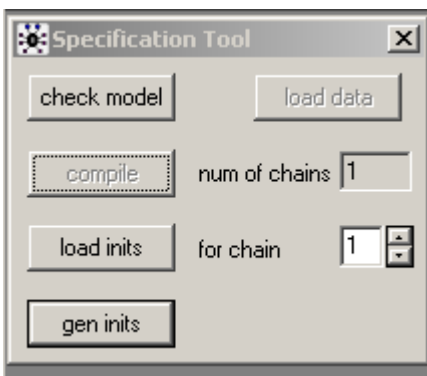
Note that models are specified as distributions.

In program code (very similar to splus code):

```
model
     {
          for(i in 1:N){
               Y[i] ~ dnorm(mu[i], tau)                    # Observation
               mu[i] <- alpha + beta * (x[i] - mean(x[]))  # System
          }
          # Distributions are set to all parameters
          sigma <- 1/sqrt(tau)
          alpha ~ dnorm(0, 1.0E-6)
          beta ~ dnorm(0, 1.0E-6)
          tau ~ dgamma(1.0E-3, 1.0E-3)
     }
```

We use the Model -> Specification menu:

Specification Tool

check model   load data

compile   num of chains 1
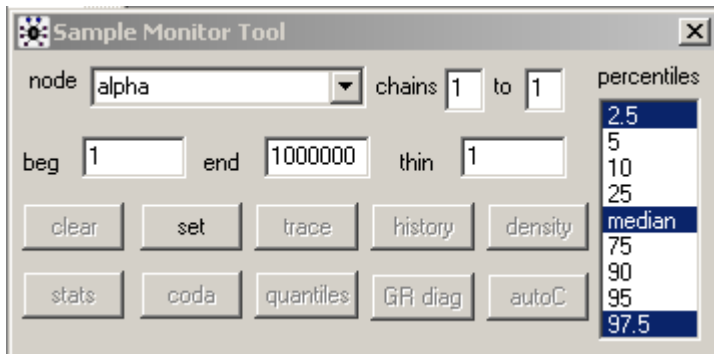
load inits   for chain   1

gen inits

to check model, load data and compile the model (in that order).

Finally, we need initial values to start the chain:

```
list(alpha = 0, beta = 0, tau = 1)
```
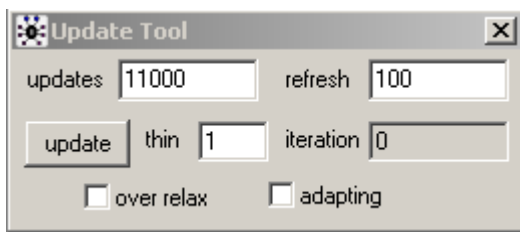
ii). We are now ready to set monitoring scheme

This is done in the Inference -> Sample Monitor Tool.

Here we decide what parameters to save and monitor:

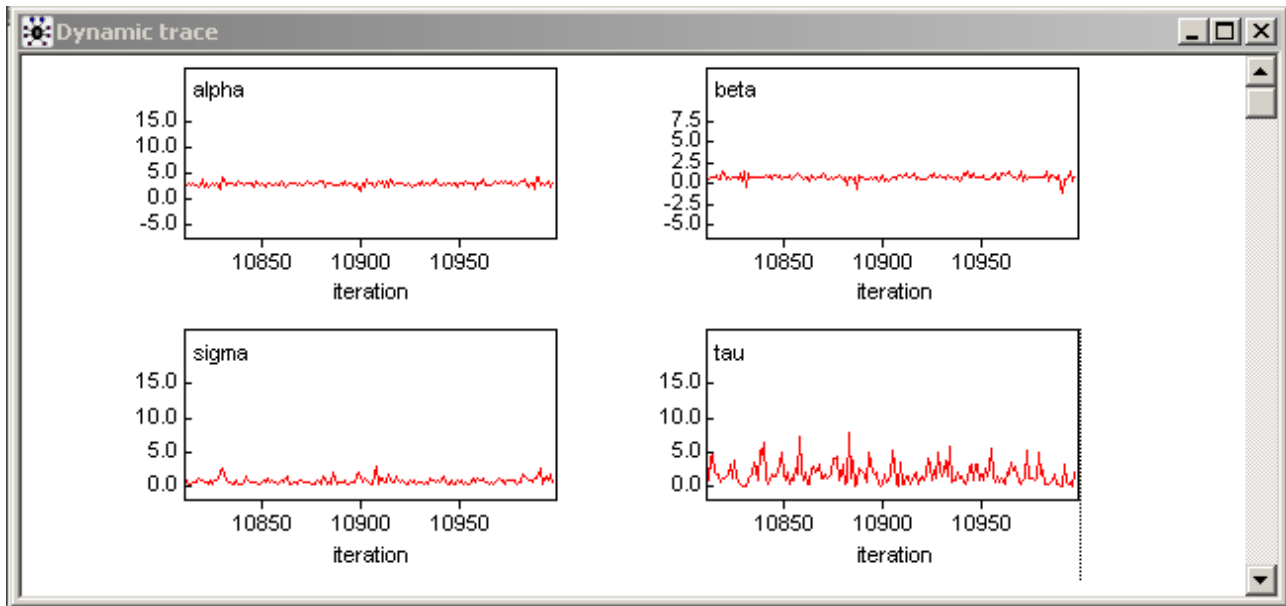(we sample alpha, beta, tau and sigma – all variables, finish with a '*').

- Then we click "trace" to see the development

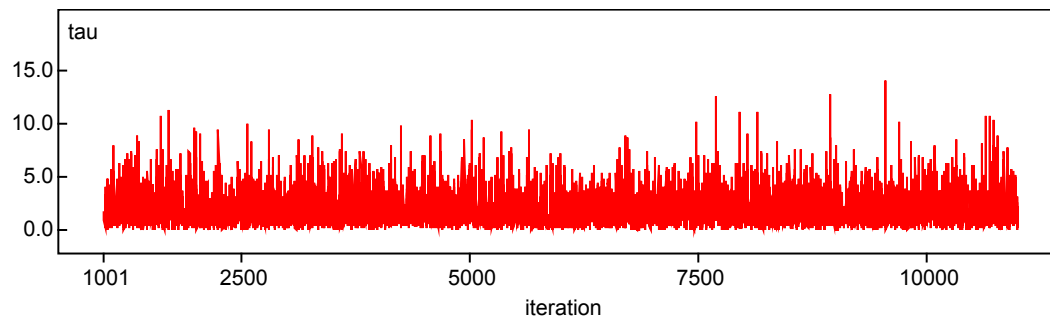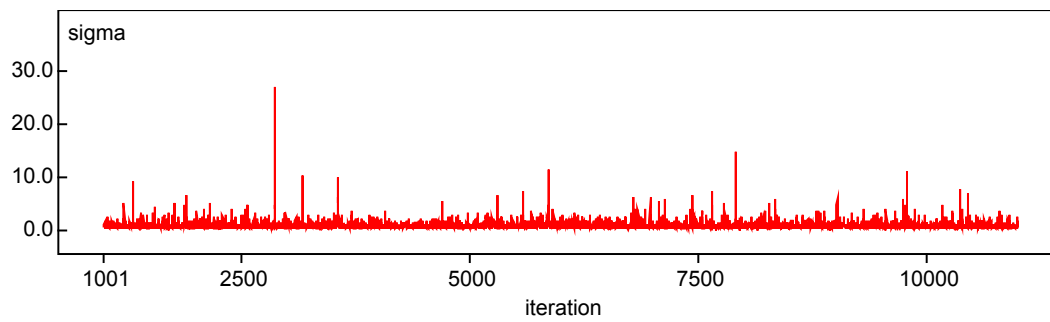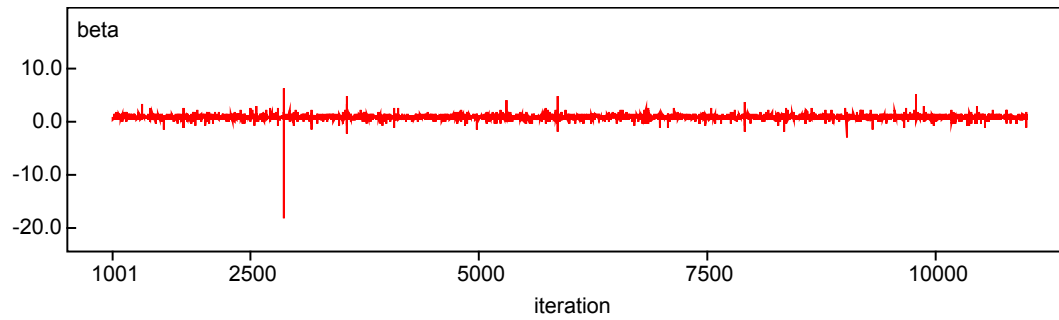iii.) We are ready to sample from the distributions specified

- We take 11.000 samples (and let the 1000 first be 'burn-in', i.e., to stabilise the values somewhat)

These are the results:

The history:

alpha

beta

sigma

tau
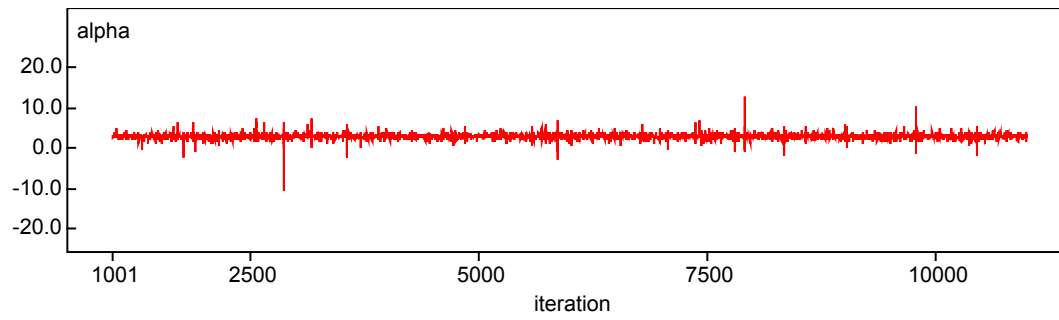
Density:
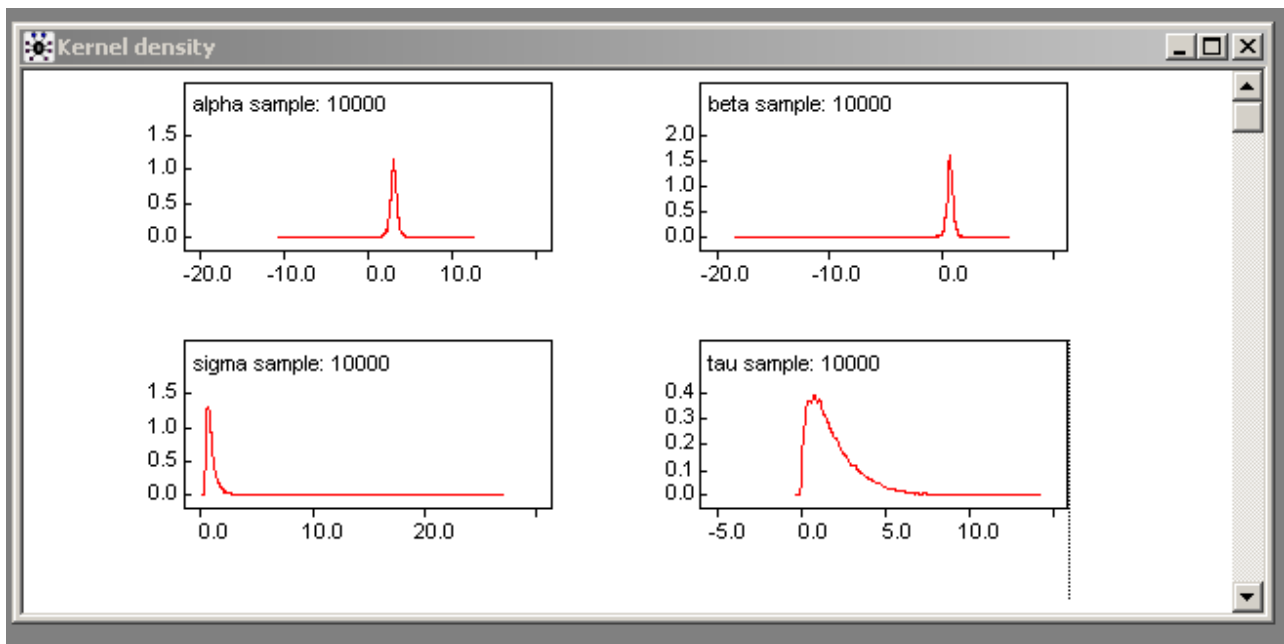
node	mean	sd	MC error	2.5%	median	97.5%	start	sample
alpha	2.99	0.5587	0.005389	1.984	2.991	3.995	1001	10000
beta	0.7967	0.4207	0.004471	0.0854	0.796	1.486	1001	10000
sigma	0.9998	0.7714	0.01375	0.4148	0.8221	2.598	1001	10000
tau	1.89	1.534	0.02352	0.1482	1.48	5.823	1001	10000

We see that the results are almost the same as for the maximum likelihood procedure of splus:

Mean of 2.99 (95% CI: 1.98 to 4.0) but with a high variability.

Beta of 0.80 (the same).

-> Some of the variance has been (incorrectly) assigned to an observation variance.

WinBUGS example II: Community dynamics in the coastal zone

Model:

$$s_t = a * s_{t-1} + r + \underline{b}^T \underline{\omega}_t + \varepsilon_t^1 \qquad \text{(ib)}$$

where $s_{t-1} = \log(S_{t-1})$. a is autoregressive parameter, b is environmental coefficients, ω is environmental data.

ε_t^1 is assumed to be normally distributed with expectation zero and some variance, say σ^2. This is equivalent to writing:

$$s_t \sim N(b * s_{t-1} + r + \underline{b}^T \underline{\omega}_t, \varepsilon_t^1) \qquad \text{(ib)}$$

This is the system equation/state equation/process equation.

We furthermore formulate an observation model:

$$S_t = \exp\{s_t\} + \varepsilon_t^2 \qquad \text{(iia)}$$

which can be reformulated as the following.

$$S_t \sim N(\exp\{s_t\}, \varepsilon_t^2) \qquad \text{(iia)}$$

where ε_t^2 is given by the estimated standard errors found in Lekve K., Boulinier T., Stenseth N.C., Gjøsæter J., Fromentin J.-M., Hines J.E. & Nichols J.D. 2002. Spatio-temporal dynamics of species richness in coastal fish communities. *Proc. R. Soc. Lond. B*, 269, 1781-1789.

In BUGS language, this is the following

 # model with environment in process model:
"C:\Kyrre\Studier\Cr\SPMOD\Oecologia\Oecol2nd\OecBug\armodel1b"

```
model
{
 s[1] ~ dnorm(0, var);
 for (i in 2:N)
  {
    muS[i] <- r + b * s[i-1]+ psi*naot1[i] +tau*temp[i]+rho*wind[i]       # Observation equation
    s[i] ~ dnorm(muS[i], var)                                            # System equation for first obs.
  }


 for (i in 1:N)
  {
   Shat[i]  ~ dnorm(ant[i], estvar[i])I(0,)                             # System equation for next obs.
   log(ant[i]) <- s[i]
  }


# Putting distributions on all parameters
r~dnorm(0,0.000001)
sd <- 1/sqrt(var)
var~dgamma(0.001,0.001)
b ~ dnorm(0,0.0001)
psi ~ dnorm(0,0.0001)
tau ~ dnorm(0,0.0001)
rho ~ dnorm(0,0.0001)
a <- 1-b
R0 <- exp(r)
for (i in 1:N)
{
S[i] <- exp(s[i])
}
meanS <- mean(S[])
}



# Data (Kragerø)
# Initialising values (two files)
```

list(var = 6, psi = 1, tau=0.5, rho=1, b = 1, r=1,

s = c(1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1))

list(var = 2, psi = 0.5, tau=1, rho=0.5, b=0.2, r=0.5,

s = c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0))


```
-> Using two chains: 110.000 iterations (time two chains: starting on 60.001):
Thinning every 10. simulation => sample size of 10.000
```

Selected results:

node	mean	sd	MC error	5.0%	median	95.0%	start	sample
R0	7.915	9.062	0.3297	1.477	5.561	21.69	60001	10000
a	0.6226	0.297	0.01095	0.1415	0.6212	1.12	60001	10000
meanS	16.66	0.1049	0.001064	16.49	16.66	16.83	60001	10000
psi	0.08219	0.1143	0.001159	-0.1044	0.08236	0.2698	60001	10000
rho	0.1977	0.1337	0.001453	-0.0187	0.1977	0.4194	60001	10000
sd	0.5879	0.07516	8.095E-4	0.4777	0.5817	0.7232	60001	10000
tau	-0.1557	0.1438	0.001589	-0.3915	-0.1567	0.08139	60001	10000

We see that

- The estimated coefficients are "well behaved": Most of them are symmetrically distributed around some value.

- There is low variability in the overall number of species.

- Several of the environmental variables are different from 0: psi on NAO, rho on wind

- Not significant autoregressive parameter

- Etc.

Assessment

A Bayesian approach

- Appealing concepts of probability etc.

- Useful for estimating both observation and system variability (much due software advances)

- Very useful for complex problems with not known distributions (*e.g.,* in DNA-analysis)

- Known likelihoods => Recursions can be explicitly determined (e.g., BTS)

- Unknown likelihoods => Simulation (BUGS)