Exercise 1: Probabilistic models and likelihood functions

Eawag Summer School in Environmental Systems Analysis

Objectives:

- Being able to derive and implement the likelihood function of a model.
- Performing maximum likelihood estimations.
- Understanding the difference between evaluating a likelihood function and sampling from ~~a likelihood function.~~ the associated probabilistic model

Toy Models:

Some exercises will use the toy models “Monod”, “Growth”, or “Survival”, which were introduced in the previous lectures. You can see their definitions for instance in the slides “Mathematical Representation and Construction of Models”.

1. Likelihood for a linear model ★

~~Analytical expression~~ Likelihood derivation

Construct the likelihood function for a simple one-dimensional linear model and additive, independent identically distributed (i.i.d.) normal errors. That is, the deterministic part of the model is given by

$$y_{det}(x, \beta, \gamma) = \beta x + \gamma \,,$$

and the probabilistic model is obtained by adding a Gaussian noise term.

Hint: Have a look at the slides “Formulation of model likelihood functions” if you need inspiration.

Likelihood evaluation

For the linear model, implement a function that returns the logarithm of the likelihood, for given model parameters and measurement data. Read the data provided in the file `model_linear.csv` and use them as “observation data” to compute the log-likelihood for two parameter sets:

$\{\beta = 2,\ \gamma = 1,\ \sigma = 1\}$ and $\{\beta = 3,\ \gamma = 2,\ \sigma = 1\}$. Which parameters are more likely to have generated this data?

Hints

R	Julia

You can read the external data using the following piece of code:

```
## read the data:
dat <- read.table("data/model_linear.csv", header=TRUE)
dat$x
dat$y

## plot the data:
plot(dat$x, dat$y, xlab='x', ylab='y')
```

Here is ~~an~~ template for the log-likelihood implementation – you still need to fill in the important bits!

```
loglikelihood.linear <- function(par, x, y){
    ## deterministic part:
    y.det <- ... # first evaluate the deterministic model output, using the para
meters par = c(beta, gamma, sigma)

    ## Calculate loglikelihood:
                                        # put your expression for log-likelihood
here
                                        # Hint: use function dnorm()

}
```

Likelihood optimization

Use an optimizer to find the parameter values that maximise the likelihood (the so called maximum likelihood estimator, MLE). Plot the resulting linear model together with the data. Does the result look reasonable?

Hints

R	Julia

You can use, for instance, use the `optim` function:

```
# Define starting parameters for optimisation (e.g. beta_init, gamma_init, and s
igma_init),
# then find maximising parameters using optim:
par.max <- optim(par = c(beta_init, gamma_init, sigma_init), fn = loglikelihood.
linear, x=x, y=y)

                                              # You can look at the result like this:
par.max$par
```

Careful: `optim` does minimisation per default!

Linear regression

Use the standard linear regression function to estimate the parameters, and compare them to the ones you found through likelihood maximisation in Exercise 1.3.

Hints

R **Julia**

The function `lm` implements linear regression.

```
## Build a linear regression model
linearModel <- lm(y ~ x, data=data)

## Inspect the model:
coef(linearModel)
summary(linearModel)
```

2. Likelihood for model "Monod" ★

Analytical expression

Construct the likelihood function for the model "Monod". Use the deterministic part of the model given in the introduction to the exercises and assume i.i.d. normal errors on the deterministic model output. The result will look very similar to the solution of Exercise 1.1!

Likelihood evaluation

Implement a function that returns the logarithm of the likelihood, for given model parameters and measurement data. Read the data provided in the file `model_monod_stoch.dat` and use them as "observation data" to compute the log-likelihood for the parameter sets $\{r_{max} = 5,\ K = 3,\ \sigma = 0.2\}$ and $\{r_{max} = 10,\ K = 4,\ \sigma = 0.2\}$. Which parameters are more likely to have generated this data?

3. Forward model simulation ★

Deterministic model simulation

Produce deterministic model outputs. Do this for the two example models:

- Model "Monod" over a concentration (C) range from 0 to 10 with the default parameter values $r_{\text{max}} = 5$ and $K = 3$.
- Model "Growth" over a time interval from 0 to 2 with default parameter values $\mu = 4$, $K = 10$, $b = 1$, $Y = 0.6$, $C_{\text{M,ini}} = 10$, and $C_{\text{S,ini}} = 50$.

Plot and interpret the results.

Hints

R **Julia**

Both deterministic models are already implemented as `model.monod` and `model.growth` in the file `models.R` .

Probabilistic model simulation

For the model "Monod", write a function that produces model outputs (i.e. samples from the probabilistic model). In other words, we want to simulate new observation data including observation noise.

- For what could that be useful?

Make the assumptions that the noise is i.i.d normal with standard deviation $\sigma = 0.2$. For the deterministic model use $r_{\text{max}} = 5$ and $K = 3$.

Use your function to simulate several thousand probabilistic model realisations (hypothetical data sets), for fixed model parameters, and plot the 10% and 90% quantiles as continuous prediction bands.

Hints

R **Julia**

For the computation of quantiles you can use the R function `quantile` .

4. Likelihood and forward simulation for the model "Survival"

Analytical expression for the likelihood

Construct the likelihood for the model "Survival". The mortality rate λ, upon multiplication with an infinitesimal time interval Δt, denotes the probability to die within this time interval, given that one is still alive at the beginning of it. If $S(t)$ denotes the probability of an individual to be still alive at time point t, this reads as

$$\frac{S(t+\Delta t)-S(t)}{S(t)} = -\lambda \Delta t\,.$$

If we let $\Delta t \to 0$ this equation turns into the differential equation

$$\dot{S}(t) = -\lambda S(t)\,.$$

Solve this equation to find the time-dependence of $S(t)$ (Hint: try an exponential ansatz or wolframalpha.com (https://www.wolframalpha.com/)). From this solution, derive the probability for an individual to die within time-interval $[t_{i-1}, t_i]$. Now, consider N independent individuals, each with the same mortality rate λ. Derive the likelihood function, for a vector of death counts $\mathbf{y} = (y_1, \dots, y_n)$, where y_i denotes the number of deaths occurring within time interval $[t_{i-1}, t_i]$, for $0 = t_0 < t_1 < \cdots < t_n$.

Hint: Look up the definition of the multinomial distribution!

Forward simulation

Write a function that simulates output from this probabilistic model, for given parameter values. Use this function to simulate several thousand model realisations, for fixed model parameters (use $N = 30$ individuals with mortality rate $\lambda = 0.2d^{-1}$ and 5 subsequent observation windows of one day each). Use a boxplot to visualize the results.

Likelihood evaluation

Implement a function that returns the logarithm of the likelihood, for given parameter values and measurement data. Check that your log-likelihood is implemented correctly by generating model outputs and computing the likelihood for several parameters, including the one that you generated the data with.

5. Sensitivity analysis (NEEDS TO BE REWRITTEN!)

For the models “Monod” and “Growth” try and compare different sensitivity analysis approaches.

Manual sensitivity analysis

Increase the model parameters by 10% and by 50%, redo the forward simulations, and try to understand the effect of the parameter change on the model results.

Local sensitivity analysis

A simple metric that measures the sensitivity of model parameters θ to model output Y is defined as

$$s_{loc} = \frac{\Delta Y}{\Delta \theta}$$

Often a relative metric is easier to interpret:

$$s_{loc} = \frac{\theta}{Y}\frac{\Delta Y}{\Delta \theta}$$

Set the ranges $\Delta\theta$ to 10% and by 50% of the parameter values and interpret the results.

Variance-based sensitivity

Conduct a variance-based regional (global) sensitivity analyses with lognormal parameter distributions based on default mean values and standard deviations of 10% and 50% of the mean and interpret the results.

Hints

R **Julia**

Use the `fast99` function implemented in the package `sensitivity`.

Comparing different sensitivity measures

Compare the results of the different sensitivity approaches. Do the rankings agree? Which one would you trust?