Exercise no 11

Corentin Lacroix 1812554

MTH 6312: Méthodes Statistiques D'Apprentissage

December 2, 2015

Exercise 1 For a single data show $\mathrm{E}\{\frac{\partial \ell(\theta)}{\partial \theta}\} = 0$ where $\ell(\theta) = \log f(y_1; \theta)$ and expectation is taken with respect to the distribution of y. Remark: use the fact that $\int_{-\infty}^{\infty} f(y_1; \theta) dy_1 = 1$.

Solution 1 With $\ell(\theta, y_1) = \log f(\theta, y_1) \forall (\theta, y_1) \in \mathcal{R}^2$, we have :

$$\frac{\partial \ell}{\partial \theta}(\theta, y_1) = \frac{1}{f(\theta, y_1)} \times \frac{\partial f}{\partial \theta}(\theta, y_1)$$

Hence,

$$\begin{aligned} \mathrm{E}\{\frac{\partial \ell(\theta)}{\partial \theta}\} &= \int_{-\infty}^{+\infty} \frac{1}{f(\theta, y_1)} \times \frac{\partial f}{\partial \theta}(\theta, y_1) \times f(\theta, y_1) dy_1 \\ &= \int_{-\infty}^{+\infty} \frac{\partial f}{\partial \theta}(\theta, y_1) dy_1 \end{aligned} \tag{1}$$

By the Leibniz rule, that we can use here as the function $\phi : y_1 \longrightarrow f(y_1, \theta)$ is derivable (we suppose it here for simplification but the proof can be established otherwise) and integrable on $\mathcal{R}$ for all $\theta \in \mathcal{R}$, we can then deduce :

$$\begin{aligned} \mathrm{E}\{\frac{\partial \ell(\theta)}{\partial \theta}\} &= \frac{\partial}{\partial \theta}(\int_{-\infty}^{+\infty} \frac{\partial f}{\partial \theta}(\theta, y_1) dy_1) \\ &= 0 \end{aligned} \tag{2}$$

As f is a density function and $\int_{-\infty}^{+\infty} \frac{\partial f}{\partial \theta}(\theta, y_1) dy_1 = 1$.

Exercise 2 Show that the same result $\mathrm{E}\{\frac{\partial \ell(\theta)}{\partial \theta}\} = 0$ still holds for the log likelihood of n data points $\ell(\theta) = \sum_{i=1}^{n} \log f(y_i; \theta)$.

Solution 2 In this case we have $\ell(\theta, y_1 \cdots, y_n) = \sum_{i=1}^{n} \log(f(\theta, y_1, \cdots, y_n))$.
So $\frac{\partial \ell}{\partial \theta}(\theta, y_1, \cdots, y_n) = \sum_{i=1}^{n} \frac{\partial f}{\partial \theta}(\theta, y_i) \times \frac{1}{f(\theta, y_i)}$

Consequently,

$$
\begin{aligned}
\mathrm{E}\{\frac{\partial \ell(\theta)}{\partial \theta}\} &= \int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty}(\sum_{i=1}^{n}\frac{\partial f}{\partial \theta}(\theta, y_i) \times \frac{1}{f(\theta, y_i)} \times \prod_{j=1}^{n} f(\theta, y_j))dy_1\cdots dy_n \\
&= \sum_{i=1}^{n}\int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty}(\frac{\partial f}{\partial \theta}(\theta, y_i) \times \frac{1}{f(\theta, y_i)} \times \prod_{j=1}^{n} f(\theta, y_j))dy_1\cdots dy_n \qquad (3) \\
&= \sum_{i=1}^{n}\int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty}((\int_{-\infty}^{+\infty}\frac{\partial f}{\partial \theta}(\theta, y_i)dy_i) \times \prod_{j\neq i}^{n} f(\theta, y_j))dy_1\cdots dy_n
\end{aligned}
$$

For the same reasons than previously, the single integral into brackets ($\int_{-\infty}^{+\infty}\frac{\partial f}{\partial \theta}(\theta, y_i)dy_i$) is equal to zero ($\forall i$), hence all the terms of that sum are zeros and we can conclude :

$$
\mathrm{E}\{\frac{\partial \ell(\theta)}{\partial \theta}\} = 0
$$

Exercise 3 Consider the linear regression `x1=rnorm(20);x2=rnorm(20);x3=rnorm(20);y=1+0.1*x1-` Fit all possible regression models (8 models) and evlauate BIC and AIC for all models.

- What is the best model according to AIC? - What is the best model according to BIC? - Make this simulation 10^6 times. How many times AIC chooses a wrong model? How many times BIC chooses a wrong model? - Increase the sample size from $n = 20$ to $n = 100$, and another time to $n = 1000$. How many times AIC chooses a wrong model. How many times BIC chooses a wrong model?

Solution 3 For this exercise, I made the following R function for testing accuracy of AIC and BIC scores for the regression problem :

```
f <- function(n_samples, n_iter, k){

    AICcount <- 0

    BICcount <- 0
```

```
for (i in 1:n_iter)
{
    c <- rep(1, n_samples)
    x1 = rnorm(n_samples)
    x2 = rnorm(n_samples)
    x3 = rnorm(n_samples)
    y=1+0.1*x1-0.1*x2+rnorm(n_samples)
    fits.lm1 <- lm(y ~ x1)
    fits.lm2 <- lm(y ~ x1 + x2)
    fits.lm3 <- lm(y ~ x1 + x3)
    fits.lm4 <- lm(y ~ x2)
    fits.lm4 <- lm(y ~ x2 + x3)
    fits.lm4 <- lm(y ~ x2)
    fits.lm5 <- lm(y ~ x2 + x3)
    fits.lm6 <- lm(y ~ x3)
    fits.lm7 <- lm(y ~ x1 + x2 + x3)
    fits.lm8 <- lm(y ~ c)
    AIC <- c(AIC(fits.lm1, k=2),  AIC(fits.lm2, k=2),  AIC(fits.lm3, k=2),
         AIC(fits.lm4, k=2),  AIC(fits.lm5, k=2),  AIC(fits.lm6, k=2),
         AIC(fits.lm7, k=2),  AIC(fits.lm8, k=2))
    BIC <- c(AIC(fits.lm1, k=log(n_samples)),  AIC(fits.lm2, k=log(n_samples)),
             AIC(fits.lm3, k=log(n_samples)),  AIC(fits.lm4, k=log(n_samples)),
             AIC(fits.lm5, k=log(n_samples)),  AIC(fits.lm6, k=log(n_samples)),
              AIC(fits.lm7, k=log(n_samples)),  AIC(fits.lm8, k=log(n_samples)))


          if (which.min(AIC) != k) AICcount <- AICcount + 1

          if (which.min(BIC) != k) BICcount <- BICcount + 1

      }

      return(c(AICcount, BICcount))

}
```

With only one test and for *n_samples* $= 20$ the best model (among the 8 considered) according to both AIC and BIC model is the last model involving only a constant.

I could repeat this process only 1000 times because of computation issues.

- For *n_samples* $= 20$, AIC score indicates a wrong best model 94.8% of the time and indicates that the model constant is the best model 35.1% of the time (which is the best score among all the models). BIC score indicates a wrong model in 97.1% of the cases and indicates that the good model is the constant model in 73.8% of the cases. For this special case, AIC may be slighly better than BIC but both are not appropriate.

- For *n_samples* $= 100$, AIC score indicates a wrong best model in 88.9% of the cases and indicates that the model constant is the best model 45.9% of the time. According to AIC score, the good model is one of the wort possible model among the 8 considered. BIC score indicates a wrong model in 98.9% of the cases and indicates that the good model is the constant model in 84.8% of the cases. BIC score is even worst than in the previous case (but it may due to the fact that the process is repeated only 1000 times), AIC gives slightly better information about accuracy of the models than in the previous case but both seem still unappropriate.

- For *n_samples* = 1000, AIC score indicates a wrong best model in only 23.1% of the cases and is this time much more powerful than with less samples. This time, the model the more often chosen is the right one. BIC score indicates a wrong model in only 49.5% of the cases. The more chosen model is also the right model. With a large nmber of samples, both AIC and BIC scores select the good model in the majority of the cases. This may be due to the fact that the approximations used to obtain an exact formulation for the AIC and BIC scores are wrong with a small number of observations and consequently they are not really appropriate in some cases. We can add here (for 1000 observations) that AIC is much more powerful than BIC for choosing better models.