18.05 Problem Set 6, Spring 2022 Solutions

Problem 1. (10 pts.) **Continuous MLE**
Suppose we have a distribution with the following pdf (called a gamma distribution)

$$f(x|a) = \frac{a^5}{(4)!} x^4 e^{-ax}.$$

Suppose we have independent data $x_1, x_2, \ldots, x_m$ drawn from this distribution. Find the maximum likelihood estimate (MLE) for a.

Suggestion: *The likelihood can be compactly written in terms of the sum S and the product P of the data.*

Solution: The likelihood for x_i is $f(x_i|a) = \frac{a^5}{4!} x_i^4 e^{-ax_i}$. So, the likelihood of the data is

$$f(\text{data}|a) = \prod_{i=1}^{m} f(x_i|a) = \frac{a^{5m}}{(4!)^m} P^4 e^{-aS},$$

where $P = \prod x_i$ (product of data) and $S = \sum x_i$ (sum of data).
So, the log likelihood is

$$l(a) = 5m\ln(a) + 4\ln(P) - aS - m\ln(4!).$$

Taking the derivative and setting it to 0, we get

$$l'(a) = \frac{5m}{a} - S = 0 \;\Rightarrow\; \boxed{\text{The MLE } \hat{a} = \frac{5m}{S}.}$$

Note: It turns out, the distribution mean is $5/a$ and $\hat{a} = 5/(S/m) = 5/\overline{x}$, where $\overline{x}$ is the data mean.

Problem 2. (25: 5,10,10 pts.) **Least squares**
In this problem we will use maximum likelihood estimates to develop Gauss' method of least squares for fitting lines to data.

Bivariate data means data of the form

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

For bivariate data the simple linear regression model assumes that, for some values of the parameters a and b, we have

$$y_i = ax_i + b + \text{ random measurement error.}$$

The model assumes the measurement errors are independent and identically distributed and follow a $N(0, \sigma^2)$ distribution. (The values x_i may or may not be random.)

In general terms, we can say that the value of x 'explains' the value of y except for some random noise. Graphically, the model says to make a scatter plot and find the line that best fits the data. This is called a simple linear regression model.

It turns out that, under some assumptions about random variation of measurement error, one way to find a "best" line is by solving a maximum likelihood problem.

The goal is to find the values of the model parameters a and b that give the MLE for this model. To guide you, we note that the model says that

$$y_i \sim N(ax_i + b, \sigma^2).$$

Also remember that you know the density function for this distribution.

(a) *For a general datum (x_1, y_1) give the likelihood and log likelihood functions (these will be functions of y_1, x_1, a, b, and σ.)*

Solution: Since $y_i \sim \mathrm{N}(ax_i + b, \sigma^2)$ the likelihood with data (x_1, y_1) is

$$f(x_1, y_1 \,|\, a, b, \sigma) = \frac{1}{\sqrt{2\pi}\,\sigma} \mathrm{e}^{-(y_1 - ax_1 - b)^2/(2\sigma^2)}.$$

The log likelihood is

$$\ln(f(x_1, y_1 \,|\, a, b, \sigma)) = -\ln(\sqrt{2\pi}\,\sigma) - \frac{(y_1 - ax_1 - b)^2}{2\sigma^2}.$$

(b) *Consider the data $(1, 8)$, $(3, 2)$, $(5, 1)$. Assume that $\sigma = 3$ is a known constant and find the maximum likelihood estimate for a and b.*

Note: *since there are* two *variables a and b, in order to find a critical point you will have to take partial derivatives and set them equal to 0. This part of the problem takes a fair amount of tedius algebra –sorry.*

Note: *We gave you a specific value of σ, to avoid the distraction of one more symbol. If you look at your calculations, you should see that the value of σ plays no role in finding the MLE for a and b. We get the same answer no matter what the value.*

Solution: The likelihood for all the data is the product of the individual likelihoods. So,

$$f((1, 8),\ (3, 2),\ (5, 1) \,|\, a, b, \sigma) = \left(\frac{1}{\sqrt{2\pi}\,\sigma}\right)^3 \mathrm{e}^{-((8-a-b)^2 + (2-3a-b)^2 + (1-5a-b)^2)/(2\sigma^2)}$$

Taking the natural log (and replacing the list of data by the word 'data') we get

$$\ln(f(\text{data} \,|\, a, b, \sigma)) = -3\ln(\sqrt{2\pi}\,\sigma) - \frac{(8-a-b)^2 + (2-3a-b)^2 + (1-5a-b)^2}{2\sigma^2}$$

Since we want to find a and b that maximize the likelihood we take the partial derivatives and set them to 0.

$$\frac{\partial \ln(f(\text{data}) \,|\, a, b, \sigma)}{\partial a} = \frac{2}{2\sigma^2}((8-a-b) + 3(2-3a-b) + 5(1-5a-b)) = 0$$
$$\frac{\partial \ln(f(\text{data}) \,|\, a, b, \sigma)}{\partial b} = \frac{2}{2\sigma^2}((8-a-b) + (2-3a-b) + (1-5a-b)) = 0$$

These are two equations in the unknowns a and b. We simplify and solve:

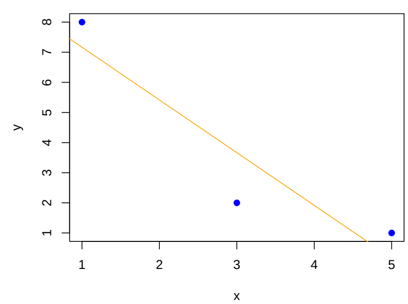
$$\begin{aligned} 35a + 9b &= 19 \\ 9a + 3b &= 11 \end{aligned} \qquad \text{which gives} \qquad a = -7/4 = -1.75; \quad b = 107/12 \approx 8.917.$$

The linear regression fit of a line to the data is $\boxed{y = ax + b = -7x/4 + 107/12.}$

(c) *Use R to plot the data and the regression line you found in part (ii) The commands `plot(x,y, pch=19)` and `abline()` will come in handy. For `abline` be careful: the parameter `a` is the intercept and `b` is the slope – exactly the opposite of our usage. Print the plot and turn it in.*

Solution: Here is the code for this plot:

```
x = c(1,3,5)
y = c(8,2,1)
a = -7/4
b = 107/12
plot(x, y, pch=19, col='blue')
#Perversely, in abline a is the intercept and b is the slope.
abline(a=b, b=a, col='orange', lwd=2)
```



Problem 3. (15: 10,5 pts.) **Estimating uniform parameters**
(a) *Suppose we have data `1.2, 2.1, 1.3, 10.5, 5` which we know is drawn indepenedently from a uniform(a, b) distribution. Give the maximum likelihood estimate for the parameters a and b.*

Hint: in this case you should not try to find the MLE by differentiating the likelihood function.

Solution: The pdf for uniform(a, b) distribution takes two values

$$f(x \mid a, b) = \begin{cases} 1/(b-a) & \text{if } x \text{ is in } [a, b] \\ 0 & \text{otherwise} \end{cases}$$

Since the likelihood is the product of the likelihoods of each data point, the likelihood function is

$$f(\text{data} \mid a, b) = \begin{cases} 1/(b-a)^5 & \text{if all data is in } [a, b] \\ 0 & \text{if not} \end{cases}$$

This is maximized when $(b - a)$ is as small as possible. Since all the data has to be in the interval $[a, b]$ we minimize $(b - a)$ by taking a = minimum of data and b = maximum of data.

Answer: $\boxed{a = 1.2,\ b = 10.5}$.

(b) *Suppose we have data $x_1, x_2, \ldots, x_n$ which we know is drawn indepenedently from a uniform(a, b) distribution. Give the maximum likelihood estimate for the parameters a and b.*

Solution: The same logic as in part (a) shows $\boxed{a = \min(x_1, \ldots, x_n) \text{ and } b = \max(x_1, \ldots, x_n)}$.