MSc. Data Science & Artificial Intelligence

Statistical Inference - Practice

Dr. Marco CORNELLI

Final assignment

By: Joris LIMONIER

joris.limonier@gmail.com

Due: January 26, 2022

Contents

1 Exercise 1

Let $(x_1, y_1), \dots (x_N, y_N)$ be observations assumed to be generated by a linear model:

$$y_i = a + bx_i + \epsilon_i \tag{LM}$$

with $a, b \in \mathbb{R}$ and $1 \leq i \leq N,\ \epsilon_i \sim \mathcal{N}(0, \sigma^2)$ independent and identically distributed (i.i.d.)

1.1 Question (a)

We compute $\mathbb{E}[y_i]$:

$$\begin{aligned}
\mathbb{E}[y_i] &= \mathbb{E}[a + bx_i + \epsilon_i] && \textit{(definition)} \\
&= \mathbb{E}[a] + \mathbb{E}[bx_i] + \mathbb{E}[\epsilon_i] && \textit{(linearity of expectation)} \\
&= a + bx_i && \textit{(}a, b, x_i \textit{ deterministic, } \epsilon_i \textit{ centered normal)}
\end{aligned}$$

We compute $\mathrm{Var}(y_i)$:

$$\begin{aligned}
\mathrm{Var}[y_i] &= \mathbb{E}\left[y_i^2\right] - \mathbb{E}[y_i]^2 \\
&= \mathbb{E}\left[(a + bx_i + \epsilon_i)^2\right] - (a + bx_i)^2 \\
&= a^2 + b^2x_i^2 + \mathbb{E}\left[\epsilon_i^2\right] && \textit{(linearity of expectation} \\
&+ 2abx_i + 2a\mathbb{E}[\epsilon_i] + 2bx_i\mathbb{E}[\epsilon_i] && \textit{and only } \epsilon_i \textit{ random)} \\
&- \left(a^2 + b^2x_i^2 + 2abx_i\right) \\
&= \mathbb{E}\left[\epsilon_i^2\right] && \textit{(}\epsilon_i \textit{ centered normal)} \\
&= \mathbb{E}\left[\epsilon_i^2\right] - \underbrace{\mathbb{E}[\epsilon_i]}_{=0}{}^2 \\
&= \mathrm{Var}(\epsilon_i) \\
&= \sigma^2
\end{aligned}$$

1.2 Question (b)

First let us note that for a given $1 \leq i \leq N$, by (LM) we have that:

$$y_i = \underbrace{a + bx_i}_{\text{deterministic}} + \underbrace{\epsilon_i}_{\mathcal{N}(0,\sigma^2)} \tag{1}$$

Moreover, we know that the Probability Density Function (PDF) of a random variable $X \sim \mathcal{N}(\mu, \sigma^2)$ is given by:

$$f_{\mu,\sigma^2}(t) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right)$$

and since for a given $1 \leq i \leq N,\ \epsilon_i \sim \mathcal{N}(0, \sigma^2)$, its PDF is given by:

$$f_{0,\sigma^2}(t) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{t^2}{2\sigma^2}\right)$$

Now applying the argument from (1) yields that for $1 \leq i \leq N$, the PDF of y_i is the following:

$$f(t;\theta) := f_{(a+bx_i),\sigma^2}(t) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\left(t-(a+bx_i)\right)^2}{2\sigma^2}\right)$$

with the parameter $\theta := (a,b) \in \mathbb{R}^2$. We write $f(t;\theta)$ (including θ), to stress the fact that we have an influence on our parameter, not over our observations.
Given this preamble, our goal is to prove the following:

$$f\left(\bigcap_{i=1}^{N} y_i;\theta\right) = \prod_{i=1}^{N} f\left(y_i;\theta\right) \tag{2}$$

We note that for a given $1 \leq i \leq N$, a, b and x_i are deterministic, so we can rewrite (2) as:

$$\begin{aligned}
f\left(\bigcap_{i=1}^{N} y_i;\theta\right) &= f\left(\bigcap_{i=1}^{N} a + bx_i + \epsilon_i;\theta\right) \\
&= f_{0,\sigma^2}\left(\bigcap_{i=1}^{N} \epsilon_i\right) && \textit{(shift of the distribution)} \\
&= \prod_{i=1}^{N} f_{0,\sigma^2}\left(\epsilon_i\right) && \textit{(ϵ_i's are i.i.d.)} \\
&= \prod_{i=1}^{N} f\left(a + bx_i + \epsilon_i\right) \\
&= \prod_{i=1}^{N} f\left(y_i;\theta\right) && (3)
\end{aligned}$$

Thus the y_i's, $1 \leq i \leq N$ are i.i.d..

1.3 Question (c)

Let $g(y_i)$ denote the PDF of y_i. We define the likelihood function of $y_i,\ 1 \leq i \leq N$ as:

$$\begin{aligned}
\mathcal{L}(\theta) &:= f(y_1, \ldots, y_N; \theta) \\
&= \prod_{i=1}^{N} f(y_i; \theta) && \textit{(y_i's are i.i.d.)}
\end{aligned}$$

We also define the log-likelihood as:

$$\ell(\theta) := \log \mathcal{L}(\theta)$$

whic is equal to:

$$\ell(\theta) = \log \prod_{i=1}^{N} f(y_i; \theta) = \sum_{i=1}^{N} \log f(y_i; \theta)$$

Thus by (3), we obtain that the log-likelihood becomes:

$$\begin{aligned}
\ell(\theta) &= \sum_{i=1}^{N} \log f(y_i; \theta) \\
&= \sum_{i=1}^{N} \log \left[\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - (a + bx_i))^2}{2\sigma^2} \right) \right] \\
&= -\frac{N}{2\sigma^2} \log \left[\frac{1}{\sqrt{2\pi\sigma^2}} \right] \sum_{i=1}^{N} (y_i - a - bx_i)^2 \qquad (4)
\end{aligned}$$

We want find $\hat{a}_{ML}$ which maximises ℓ with respect to a. To do so, we take the partial derivative with respect to a and set it to 0. Therefore we get:

$$\begin{aligned}
&\frac{\partial \ell}{\partial a} = 0 \\
\implies &\frac{\partial}{\partial a}\left[-\frac{N}{2\sigma^2}\log\left[\frac{1}{\sqrt{2\pi\sigma^2}}\right]\sum_{i=1}^{N}\left(y_i - bx_i - \hat{a}_{ML}\right)^2\right] = 0 \\
\implies &-\frac{N}{2\sigma^2}\log\left[\frac{1}{\sqrt{2\pi\sigma^2}}\right]\sum_{i=1}^{N}\frac{\partial}{\partial a}\left[\left(y_i - bx_i - \hat{a}_{ML}\right)^2\right] = 0 \\
\implies &\sum_{i=1}^{N}\frac{\partial}{\partial a}\left[\left(y_i - bx_i - \hat{a}_{ML}\right)^2\right] = 0 \\
\implies &\sum_{i=1}^{N}-2\left(y_i - bx_i - \hat{a}_{ML}\right) = 0 \\
\implies &\sum_{i=1}^{N}\left(y_i - bx_i - \hat{a}_{ML}\right) = 0 \\
\implies &-N\hat{a}_{ML} + \sum_{i=1}^{N} y_i - bx_i = 0 \\
\implies &\hat{a}_{ML} = \frac{1}{N}\sum_{i=1}^{N} y_i - bx_i
\end{aligned}$$

where we used the linearity of differentiation.
Defining $\bar{x}, \bar{y}$ as follows:

$$\begin{aligned}
\bar{x} &:= \frac{1}{N}\sum_{i=1}^{N} x_i \\
\bar{y} &:= \frac{1}{N}\sum_{i=1}^{N} y_i
\end{aligned}$$

we can rewrite $\hat{a}_{ML}$ as:

$$\hat{a}_{ML} = \bar{y} - b\bar{x} \tag{5}$$

1.4 Question (d)

Now we want to find $\hat{b}_{ML}$. We proceed similarily, that is we differentiate ℓ with respect to b and equate it to 0:

$$
\begin{aligned}
&\frac{\partial \ell}{\partial b} = 0 \\
\implies &\frac{\partial}{\partial b}\left[-\frac{N}{2\sigma^2}\log\left[\frac{1}{\sqrt{2\pi\sigma^2}}\right]\sum_{i=1}^{N}\left(y_i - a - \hat{b}_{ML}x_i\right)^2\right] = 0 \\
\implies &-\frac{N}{2\sigma^2}\log\left[\frac{1}{\sqrt{2\pi\sigma^2}}\right]\sum_{i=1}^{N}\frac{\partial}{\partial b}\left[\left(y_i - a - \hat{b}_{ML}x_i\right)^2\right] = 0 \\
\implies &\sum_{i=1}^{N}\frac{\partial}{\partial b}\left[\left(y_i - a - \hat{b}_{ML}x_i\right)^2\right] = 0 \\
\implies &\sum_{i=1}^{N} -2x_i\left(y_i - a - \hat{b}_{ML}x_i\right) = 0 \\
\implies &\sum_{i=1}^{N} x_i\left(y_i - a\right) - \sum_{i=1}^{N}\hat{b}_{ML}x_i^2 = 0 \\
\implies &\sum_{i=1}^{N} x_i\left(y_i - a\right) = \hat{b}_{ML}\sum_{i=1}^{N}x_i^2 \\
\implies &\hat{b}_{ML} = \frac{\sum_{i=1}^{N} x_i\left(y_i - a\right)}{\sum_{j=1}^{N} x_j^2}
\end{aligned}
$$

where we used the linearity of differentiation.
We now plug in the value of $\hat{a}_{ML}$ found in (5), which gives us the following system:

$$
\begin{aligned}
&\begin{cases}
\hat{a}_{ML} = \bar{y} - \hat{b}_{ML}\bar{x} \\
\hat{b}_{ML} = \frac{1}{\sum_{j=1}^{N} x_j^2} \sum_{i=1}^{N} x_i \left(y_i - \hat{a}_{ML}\right)
\end{cases} \\
\implies &\begin{cases}
\hat{a}_{ML} = \bar{y} - \hat{b}_{ML}\bar{x} \\
\hat{b}_{ML} = \frac{1}{\sum_{j=1}^{N} x_j^2} \sum_{i=1}^{N} x_i \left(y_i - (\bar{y} - \hat{b}_{ML}\bar{x})\right)
\end{cases} \\
\implies &\begin{cases}
\hat{a}_{ML} = \bar{y} - \hat{b}_{ML}\bar{x} \\
\hat{b}_{ML} \sum_{i=1}^{N} x_i^2 = \sum_{i=1}^{N} x_i y_i - x_i \bar{y} + \sum_{i=1}^{N} \hat{b}_{ML}\bar{x} x_i
\end{cases} \\
\implies &\begin{cases}
\hat{a}_{ML} = \bar{y} - \hat{b}_{ML}\bar{x} \\
\hat{b}_{ML} \left[\sum_{i=1}^{N} x_i^2 - \bar{x} x_i\right] = \sum_{i=1}^{N} x_i y_i - x_i \bar{y}
\end{cases} \\
\implies &\begin{cases}
\hat{a}_{ML} = \bar{y} - \hat{b}_{ML}\bar{x} \\
\hat{b}_{ML} = \left[\sum_{i=1}^{N} x_i (y_i - \bar{y})\right] / \left[\sum_{i=1}^{N} x_i (x_i - \bar{x})\right]
\end{cases}
\end{aligned}
$$

Now, let x and y be as follows:

$$
x := \begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix} \in \mathbb{R}^n
$$

$$
y := \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} \in \mathbb{R}^n
$$

and let $\langle \cdot, \cdot \rangle$ denote the classical dot product.
Then, we rewrite our previous system and inject $\hat{b}_{ML}$ into the equation of $\hat{a}_{ML}$ to solve it.

$$
\begin{aligned}
&\begin{cases}
\hat{a}_{ML} = \bar{y} - \hat{b}_{ML}\bar{x} \\
\hat{b}_{ML} = \langle x, y - \bar{y} \rangle / \langle x, x - \bar{x} \rangle
\end{cases} \\
\implies &\begin{cases}
\hat{a}_{ML} = \bar{y} - \left[\bar{x} \langle x, y - \bar{y} \rangle / \langle x, x - \bar{x} \rangle\right] \\
\hat{b}_{ML} = \langle x, y - \bar{y} \rangle / \langle x, x - \bar{x} \rangle
\end{cases}
\end{aligned}
$$

1.5 Question (e)

We know that the Ordinary Least Squares (OLS) estimates are found by minimising the Residual Sum of Squares, which is given by:

$$R_{SS}(a,b) := \sum_{i=1}^{N} (y_i - a - bx_i)^2 = \sum_{i=1}^{N} \epsilon_i^2$$

In other words, the OLS estimates are given by:

$$(\hat{a}_{OLS}, \hat{b}_{OLS}) = \arg \min_{(a,b)\in\mathbb{R}^2} R_{SS}(a,b)$$

We have that the Maximum Likelihood estimators $(\hat{a}_{ML}, \hat{b}_{ML})$ maximise our log-likelihood function ℓ, which means:

$$\begin{aligned}
&\left(\hat{a}_{ML}, \hat{b}_{ML}\right) = \arg \max_{(a,b)\in\mathbb{R}^2} \ell(\theta) \\
\implies &\left(\hat{a}_{ML}, \hat{b}_{ML}\right) = \arg \max_{(a,b)\in\mathbb{R}^2} -\frac{N}{2\sigma^2} \log\left[\frac{1}{\sqrt{2\pi\sigma^2}}\right] \sum_{i=1}^{N} (y_i - a - bx_i)^2 \\
\implies &\left(\hat{a}_{ML}, \hat{b}_{ML}\right) = \arg \min_{(a,b)\in\mathbb{R}^2} \underbrace{\frac{N}{2\sigma^2} \log\left[\frac{1}{\sqrt{2\pi\sigma^2}}\right]}_{\text{doesn't depend on } (a,b)} \sum_{i=1}^{N} (y_i - a - bx_i)^2 \\
\implies &\left(\hat{a}_{ML}, \hat{b}_{ML}\right) = \arg \min_{(a,b)\in\mathbb{R}^2} \sum_{i=1}^{N} (y_i - a - bx_i)^2 \\
\implies &\left(\hat{a}_{ML}, \hat{b}_{ML}\right) = \arg \min_{(a,b)\in\mathbb{R}^2} R_{SS}(a,b) \\
\implies &\left(\hat{a}_{ML}, \hat{b}_{ML}\right) = \left(\hat{a}_{OLS}, \hat{b}_{OLS}\right)
\end{aligned}$$

We get that for $1 \leq i \leq N,\ \epsilon_i \sim \mathcal{N}(0, \sigma^2)$, the maximum-likelihood estimates and the OLS are equal.

2 Exercise 2

2.1 Question (a)

We compute $\mathbb{E}\left[\hat{\beta}_{OLS}\right]$

$$
\begin{aligned}
\mathbb{E}\left[\hat{\beta}_{OLS}\right] &= \mathbb{E}\left[\left(X^T X\right)^{-1} X^T Y\right] \\
&= \mathbb{E}\left[\left(X^T X\right)^{-1} X^T \left(X\beta + \epsilon\right)\right] \\
&= \mathbb{E}\left[\left(X^T X\right)^{-1} X^T X\beta + \left(X^T X\right)^{-1} X^T \epsilon\right] \\
&= \mathbb{E}\left[\beta\right] + \mathbb{E}\left[\underbrace{\left(X^T X\right)^{-1} X^T}_{\text{deterministic}} \epsilon\right] && \textit{(linearity of expectation)} \\
&= \mathbb{E}\left[\beta\right] + \left(X^T X\right)^{-1} X^T \underbrace{\mathbb{E}\left[\epsilon\right]}_{=0} \\
&= \mathbb{E}\left[\beta\right] && (\epsilon \sim \mathcal{N}(0, \sigma^2 I_N)) \\
&= \beta && (\beta \textit{ deterministic})
\end{aligned}
$$

We have $\mathbb{E}\left[\hat{\beta}_{OLS}\right] = \beta$, therefore $\hat{\beta}_{OLS}$ is an unbiased estimator.

2.2 Question (b)

We know that:

$$
\operatorname{Var}\left[\hat{\beta}_{OLS}\right] = \mathbb{E}\left[\left(\hat{\beta}_{OLS} - \beta\right)\left(\hat{\beta}_{OLS} - \beta\right)^T\right]
$$

Let us first exaluate $\hat{\beta}_{OLS} - \beta$:

$$
\begin{aligned}
\hat{\beta}_{OLS} - \beta &= \left(X^T X\right)^{-1} X^T Y - \beta \\
&= \left(X^T X\right)^{-1} X^T \left(X\beta + \epsilon\right) - \beta \\
&= \beta + \left(X^T X\right)^{-1} X^T \epsilon - \beta \\
&= \left(X^T X\right)^{-1} X^T \epsilon
\end{aligned}
$$

We now compute $\text{Var}\left[\hat{\beta}_{OLS}\right]$:

$$\begin{aligned}
\text{Var}\left[\hat{\beta}_{OLS}\right] &= \mathbb{E}\left[\left(\hat{\beta}_{OLS}-\beta\right)\left(\hat{\beta}_{OLS}-\beta\right)^T\right] \\
&= \mathbb{E}\left[\left(\left(X^TX\right)^{-1}X^T\epsilon\right)\left(\left(X^TX\right)^{-1}X^T\epsilon\right)^T\right] \\
&= \mathbb{E}\left[\left(X^TX\right)^{-1}X^T\epsilon\epsilon^TX\left(\left(X^TX\right)^{-1}\right)^T\right] \\
&= \mathbb{E}\left[\left(X^TX\right)^{-1}X^T\epsilon\epsilon^TX\left(\left(X^TX\right)^T\right)^{-1}\right] \\
&= \mathbb{E}\left[\left(X^TX\right)^{-1}X^T\epsilon\epsilon^TX\left(X^TX\right)^{-1}\right] \\
&= \left(X^TX\right)^{-1}X^T\mathbb{E}\left[\epsilon\epsilon^T\right]X\left(X^TX\right)^{-1} \\
&= \left(X^TX\right)^{-1}X^T\sigma^2 I_N X\left(X^TX\right)^{-1} && \textit{(}\ \mathbb{E}[\epsilon\epsilon^T]=\sigma^2 I_N,\ \textit{since} \\
& && \epsilon_i \sim \mathcal{N}(0,\sigma^2)\ \textit{i.i.d.)} \\
&= \sigma^2\left(X^TX\right)^{-1}X^TX\left(X^TX\right)^{-1} && \textit{(}\sigma^2\in\mathbb{R}, \\
& && \textit{therefore commutes)} \\
&= \sigma^2\left(X^TX\right)^{-1}
\end{aligned}$$

where we used that:

- The transpose of a product is the product of the transposed factors in reverse order
- The transpose of the inverse is the inverse of the transpose.