Applied Statistics (ECS764P) - Lab 4

Fredrik Dahlqvist

6 Dec 2023

1 Theory

1. Consider the measure $\mathbb{P}$ on $\{1,2,3\} \times \{1,2,3\}$ defined by

$$\begin{array}{lll} \mathbb{P}(1,1) = 1/10 & \mathbb{P}(1,2) = 2/10 & \mathbb{P}(1,3) = 1/10 \\ \mathbb{P}(2,1) = 1/10 & \mathbb{P}(2,2) = 1/10 & \mathbb{P}(2,3) = 2/10 \\ \mathbb{P}(3,1) = 1/10 & \mathbb{P}(3,2) = 0 & \mathbb{P}(3,3) = 1/10 \end{array}$$

(where I've written $\mathbb{P}(x,y)$ for $\mathbb{P}(\{(x,y)\})$ in order to keep things readable.)

(a) Is $\mathbb{P}$ a probability measure?

(b) Prove that $\mathbb{P}$ cannot be written as a product measure. *Hint: prove it by contradiction.*

(c) Compute the two marginals of $\mathbb{P}$.

(d) Compute the covariance and correlation of $\mathbb{P}$.

Answer:

(a) $\mathbb{P}$ is a probability measure since $\sum_{x,y} \mathbb{P}(x,y) = 1$.

(b) Suppose that there exists $\mathbb{P}_1$ and $\mathbb{P}_2$ on $\{1,2,3\}$ such that

$$\mathbb{P} = \mathbb{P}_1 \otimes \mathbb{P}_2$$

If this was the case, we would have $\mathbb{P}(3,2) = (\mathbb{P}_1 \otimes \mathbb{P}_2)(3,2) = \mathbb{P}_1(3)\mathbb{P}_2(2) = 0$. This is only possible if either $\mathbb{P}_1(3) = 0$ or $\mathbb{P}_2(2) = 0$. Assume first that $\mathbb{P}_1(3) = 0$, then we would have $\mathbb{P}(3,1) = \mathbb{P}_1(3)\mathbb{P}_2(1) = 0$ too, but in fact $\mathbb{P}(3,1) = 1/10$ so we get a contradiction. Suppose instead that $\mathbb{P}_2(2) = 0$, this would mean that $\mathbb{P}(2,2) = \mathbb{P}_1(2)\mathbb{P}_2(2) = 0$, but we have $\mathbb{P}(2,2) = 1/10$. Again a contradiction. This means that there cannot exist $\mathbb{P}_1, \mathbb{P}_2$ such that $\mathbb{P} = \mathbb{P}_1 \otimes \mathbb{P}_2$.

(c) By definition

$$\begin{aligned} (\pi_1)_*\mathbb{P}(\{1\}) &= \mathbb{P}(\{(1,y) \mid y \in \{1,2,3\}\}) = \mathbb{P}(1,1) + \mathbb{P}(1,2) + \mathbb{P}(1,3) = 4/10 \\ (\pi_1)_*\mathbb{P}(\{2\}) &= \mathbb{P}(\{(2,y) \mid y \in \{1,2,3\}\}) = \mathbb{P}(2,1) + \mathbb{P}(2,2) + \mathbb{P}(2,3) = 4/10 \\ (\pi_1)_*\mathbb{P}(\{3\}) &= \mathbb{P}(\{(3,y) \mid y \in \{1,2,3\}\}) = \mathbb{P}(3,1) + \mathbb{P}(3,2) + \mathbb{P}(3,3) = 2/10 \end{aligned}$$

And similarly,

$$\begin{aligned} (\pi_2)_*\mathbb{P}(\{1\}) &= \mathbb{P}(\{(x,1) \mid x \in \{1,2,3\}\}) = \mathbb{P}(1,1) + \mathbb{P}(2,1) + \mathbb{P}(3,1) = 3/10 \\ (\pi_2)_*\mathbb{P}(\{2\}) &= \mathbb{P}(\{(x,2) \mid x \in \{1,2,3\}\}) = \mathbb{P}(1,2) + \mathbb{P}(2,2) + \mathbb{P}(3,2) = 3/10 \\ (\pi_2)_*\mathbb{P}(\{3\}) &= \mathbb{P}(\{(x,3) \mid x \in \{1,2,3\}\}) = \mathbb{P}(1,3) + \mathbb{P}(2,3) + \mathbb{P}(3,3) = 4/10 \end{aligned}$$

(d) Let's start by computing the mean and variances of the marginals. We get

$$\begin{aligned} \mu_1 &= 1 \cdot \frac{4}{10} + 2 \cdot \frac{4}{10} + 3 \cdot \frac{2}{10} = \frac{18}{10} \\ \mu_2 &= 1 \cdot \frac{3}{10} + 2 \cdot \frac{3}{10} + 3 \cdot \frac{4}{10} = \frac{21}{10} \\ \sigma_1^2 &= \left(1 - \frac{18}{10}\right)^2 \cdot \frac{4}{10} + \left(2 - \frac{18}{10}\right)^2 \cdot \frac{4}{10} + \left(3 - \frac{18}{10}\right)^2 \cdot \frac{2}{10} = \frac{560}{1000} \\ \sigma_2^2 &= \left(1 - \frac{21}{10}\right)^2 \cdot \frac{3}{10} + \left(2 - \frac{21}{10}\right)^2 \cdot \frac{3}{10} + \left(3 - \frac{21}{10}\right)^2 \cdot \frac{4}{10} = \frac{690}{1000} \end{aligned}$$

We can now compute the covariance and the correlation.

$$
\begin{aligned}
Cov(\mathbb{P}) =& \left(1-\frac{18}{10}\right)\left(1-\frac{21}{10}\right)\frac{1}{10}+\left(1-\frac{18}{10}\right)\left(2-\frac{21}{10}\right)\frac{2}{10}+\left(1-\frac{18}{10}\right)\left(3-\frac{21}{10}\right)\frac{1}{10}+\\
&\left(2-\frac{18}{10}\right)\left(1-\frac{21}{10}\right)\frac{1}{10}+\left(2-\frac{18}{10}\right)\left(2-\frac{21}{10}\right)\frac{1}{10}+\left(2-\frac{18}{10}\right)\left(3-\frac{21}{10}\right)\frac{2}{10}+\\
&\left(3-\frac{18}{10}\right)\left(1-\frac{21}{10}\right)\frac{1}{10}+\left(3-\frac{18}{10}\right)\left(3-\frac{21}{10}\right)\frac{1}{10}\\
=&\frac{20}{1000}\\
Corr(\mathbb{P}) =& \frac{Cov(\mathbb{P})}{\sigma_1\sigma_2} = \frac{20}{\sqrt{560*690}} \approx 0.03
\end{aligned}
$$

2 Practice

You can assume that `numpy`, `matplotlib`, `scipy` *and* `pandas-datareader` *are installed on the machine of the person who will run and mark your notebook. There is no need to force an install with the* `!` *command. For textual answers please use a markdown cell.*

1. **(5 marks)** You will first download the world GDP data from the World Bank using `pandas_datareader`. The following code will download and plot the entire world GDP time series. Do NOT make any local copies of your data!

```
    from pandas_datareader import wb
    import matplotlib.pyplot as plt
    import numpy as np

    gdp_data = wb.downloa\P(indicator='NY.GDP.MKTP.CD', country='WLD',
start='1960', end='2021')
    time = np.arange(1960,2022)
    gdp = gdp_data.iloc[:,0].astype(float).to_numpy()
    # Data is returned in inverse chronological order, so reverse order
    gdp = np.flip(gdp)
    # Plot world GDP data against time
    plt.plot(time,gdp,label='US GDP')
    plt.legen\P()
    plt.show()
```

(you can ignore the warning about the code 'WLD'). You will try to estimate the long-term annual growth rate of the world using a regression.

(a) If the growth rate was a constant r, then the world's GDP would grow as

$$GDP_k = GDP_0(1+r)^k$$

where k is the number of years since 1960 and GDP_0 is the world's GDP in 1960. This is clearly not a linear relationship between time (k, in years) and GDP. However, we can get a linear relationship by applying a simple transformation $f(-)$ on both side of the equation. What is this transformation? (*Hint: we used this transformation in the context of MLE, it turns products into sums.*)

(b) Apply this transformation $f(-)$ to the GDP data, and perform a regression against the time variable. On the same plot, display your regression line, a scatter-plot of the (transformed) data points, and your R^2 value.

(c) Compute the residuals of your regression (i.e. the difference between the model and the observations), and print their mean and their standard deviation $\hat{\sigma}$. Perform a KS-test to determine whether we can reject the null hypothesis that the residuals are sampled from a normal distribution with mean 0 and standard deviation $\hat{\sigma}$. Take $\alpha = 99\%$.

(d) You will now apply the inverse of the transformation $f(-)$ to your linear model in order to get a non-linear model for the GDP. On the same plot, display your (non-linear) model and a scatter-plot of the (original) data points.

(e) What is the relationship between the slope of the regression and the long-term growth rate of the world GDP? Compute the long-term growth rate of the world GDP.

(f) What do you observe since approximately 2015?

2. **(5 marks)** In this question you will study the distribution of the slope and intercept parameters of a linear model. Consider the following model

$$y_i = ax_i + b + \varepsilon_i \qquad \text{where} \qquad a = \frac{1}{2}, b = 2, \varepsilon_i \sim \text{Normal}\left(0, \frac{1}{5}\right), 1 \leq i \leq N \tag{1}$$

For the purpose of this exercise you will take $N = 200$ and generate the x_is by

$$\texttt{x} = \texttt{np.linspace}(-5, 5, 200)$$

(a) Generate 10000 sets of error vectors ε_i and use them to perform 10000 linear regression of the N-dimensional vectors (y_i) against (x_i), where y_i is given by (1).

(b) Collect the slopes and the intercepts of these 10000 linear regressions and plot their histograms against their respective theoretical densities given in the lecture. What do you observe?

(c) For each of the 10000 regression, compute the test statistic for the slope and for the intercept (given in the lecture) and plot their histograms against their theoretical density (also given in the lecture). What do you observe?

(d) Take the last of your regressions and perform the following two tests with $\alpha = 99\%$ (you may use either p-values or critical regions but make sure you think about whether this is a one-sided or two-sided test).

$$\text{First test: } H_0 : a = \frac{1}{2} \qquad \text{(assuming } b = 2\text{)}$$
$$\text{Second test: } H_0 : b = 2 \qquad \text{(assuming } a = \frac{1}{2}\text{)}$$

(e) Change the model to

$$y_i = ax_i + b + \varepsilon_i \qquad \text{where} \qquad a = \frac{1}{2}, b = 2, \varepsilon_i \sim \text{Cauchy}\left(0, \frac{1}{5}\right), 1 \leq i \leq N \tag{2}$$

Perform another 10000 regressions based on this model. Collect the slopes and intercepts of these regressions as well as the associated statistics. Plot their histograms. What do you observe?