Linear regression approach to a bidimensional normally distributed random vector

Theoretical discussion and empirical application

Adrià Marín Salvador (1456429)
Albert Rodas Barceló (1418925)
Marc Seguí Coll (1465613)
Universitat Autònoma de Barcelona

June 2020

Abstract

Let (X, Y) be a bidemensional normally distributed random vector. It is shown that the conditional distribution $(Y|X = x)$ is normally distributed with mean linearly dependant on x. Hence, one can find a linear regression to estimate the value of Y given that $X = x$.

This has been studied in a particular case, relating the height and the weight of a population sample. The coefficients of the linear regression have been calculated both theoretically - (α_0, β_0) - and experimentally by the method of least squares - $(\hat{\alpha}, \hat{\beta})$ -, and a hypothesis testing has been performed to ensure the validity of our theoretical results. With a 5% significance level, it has been found that only in about 5.11% of all simulations the hypothesis $\hat{\alpha} = \alpha_0$ has been rejected, and similarly for the hypothesis $\hat{\beta} = \beta_0$. Also, in the vast majority of the simulations in which one of both hypothesis was rejected, both hypothesis were rejected.

Finally, a confidence and a prediction interval have been computed for the weight of an individual with a particular height.

1 Introduction

Definition 1. *Let $Z_1, Z_2 \approx N(0,1)$. Let $\vec{\mu} = (\mu_1, \mu_2) \in \mathbb{R}^2$ and $A \in M_2(\mathbb{R})$. Then, the random vector $(X,Y) = \vec{\mu} + A(Z_1, Z_2)^t$ is said to follow a bidimensional normal distribution, with mean vector $E[(X,Y)] = \vec{\mu}$ and covariance matrix $\Sigma = AA^t$. We write $(X,Y) \approx N(\vec{\mu}, \Sigma)$.*

Theorem 1. *Let (X,Y) be a bidimensional normally distributed random vector such that A is non-singular. Then, (X,Y) has probability density function* [1]

$$f_{(X,Y)}(x,y) = \frac{1}{2\pi|\Sigma|^{1/2}} \exp\Big\{ -\frac{1}{2}(x-\mu_1, y-\mu_2)\Sigma^{-1}(x-\mu_1, y-\mu_2)^t \Big\}$$

Remark. *Unless otherwise stated, it will be assumed that the matrix A defining all bidimensional normally distributed random vectors is non-singular.*

Theorem 2. *Let $(X,Y) \approx N_2(\vec{\mu}, \Sigma)$. Then, X and Y are independent if and only if $\rho = \mathrm{Corr}(X,Y) = 0$.*

Proof. It is known that $(X,Y) \approx N_2(\vec{\mu}, \Sigma)$ if and only if $(X,Y) = \vec{\mu} + A(Z_1, Z_2)$, where

$$AA^t = \Sigma := \begin{pmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{pmatrix}$$

and $Z_1 \approx N(0,1)$ and $Z_2 \approx N(0,1)$ are independent.

Let us define $A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$ and $\vec{\mu} = (\mu_1, \mu_2)$. Hence, one obtains

$$\begin{cases} X = \mu_1 + a_{11}Z_1 + a_{12}Z_2 \\ Y = \mu_2 + a_{21}Z_1 + a_{22}Z_2 \end{cases}$$

from which we can easily compute $E[X] = \mu_1$ and $E[Y] = \mu_2$, using the linearity of the expected value operator and the fact that Z_i have expected value 0. Hence, $E[X]E[Y] = \mu_1\mu_2$.

Now, let us compute $XY = \mu_1\mu_2 + a_{11}a_{21}Z_1^2 + a_{12}a_{22}Z_2^2 + \mu_1 a_{21}Z_1 + \mu_1 a_{22}Z_2 + \mu_2 a_{11}Z_1 + a_{11}a_{22}Z_1Z_2 + \mu_2 a_{12}Z_2 + a_{12}a_{21}Z_1Z_2$. Hence, using again the linearity of the expected value operator and the fact that Z_1 and Z_2 are independent, $E[XY] = \mu_1\mu_2 + a_{11}a_{21} + a_{12}a_{22}$, where we have made use of the fact that $E[Z_i^2] = V[Z_i] - E[Z_i]^2 = 1$, for $i = 1, 2$.

One the other hand, one can compute

$$\Sigma := \begin{pmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{pmatrix} = AA^t = \begin{pmatrix} a_{11}^2 + a_{12}^2 & a_{11}a_{21} + a_{12}a_{22} \\ a_{11}a_{21} + a_{12}a_{22} & a_{21}^2 + a_{22}^2 \end{pmatrix}$$

from which $\sigma_{xy} = a_{11}a_{21} + a_{12}a_{22}$. One also gets that σ_x and σ_y are non-zero. Indeed, otherwise, $|A| = 0$, which is not true by hypothesis.

Hence, using that $\sigma_{xy} = \rho\sigma_x\sigma_y$,

$$\rho = 0 \iff \sigma_{xy} = 0 \iff a_{11}a_{21} + a_{12}a_{22} = 0 \iff \mu_1\mu_2 + a_{11}a_{21} + a_{12}a_{22} = \mu_1\mu_2 \iff$$
$$\iff E[XY] = E[X]E[Y] \iff X \text{ and } Y \text{ are independent}$$

□

Lemma 1. *Let $(X,Y) \approx N(\vec{\mu}, \Sigma)$. Then, $X \approx N(\mu_1, \sigma_x^2)$ and $Y \approx N(\mu_2, \sigma_y^2)$, where $\vec{\mu} = (\mu_1, \mu_2)$ and*

$$\Sigma = \begin{pmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{pmatrix}$$

Proof. Let $A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$ such that $AA^t = \Sigma$. Then, as seen in the previous proof, $X = \mu_1 + a_{11}Z_1 + a_{12}Z_2$.

Hence, $X \approx N(\mu_1, a_{11}^2 + a_{12}^2) = N(\mu_1, \sigma_x^2)$ [2], since $a_{11}^2 + a_{12}^2 = \sigma_x^2$, as seen in the previous proof.

The proof for Y is analogous. From now on, we will use indistinctly $\mu_X = \mu_1$ and $\mu_Y = \mu_2$. $\square$

Theorem 3. *Let $(X,Y) \approx N(\vec{\mu}, \Sigma)$. Let $(Y|X=x)$ be the conditional distribution of Y given that $X = x$. Then*

$$(Y|X=x) \approx N\Big(\mu_2 + \frac{\sigma_y}{\sigma_x}\rho(x-\mu_1), (1-\rho^2)\sigma_y^2\Big)$$

Proof. Let $(X,Y) \approx N(\vec{\mu}, \Sigma)$. Then, by Theorem 1, (X,Y) has the following probability density function

$$f_{(X,Y)}(x,y) = \frac{1}{2\pi|\Sigma|^{1/2}} \exp\Big\{ -\frac{1}{2}(x-\mu_1, y-\mu_2)\Sigma^{-1}(x-\mu_1, y-\mu_2)^t \Big\}$$

where $\Sigma = \begin{pmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{pmatrix} = \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix}$, having defined $\sigma_{xy} = \rho\sigma_x\sigma_y$.

Hence, one can easily compute $\Sigma^{-1} = \frac{1}{|\Sigma|}\begin{pmatrix} \sigma_y^2 & -\rho\sigma_x\sigma_y \\ -\rho\sigma_x\sigma_y & \sigma_x^2 \end{pmatrix}$

Let us define

$$\begin{aligned} Q(x,y) :=& (x-\mu_1, y-\mu_2)\Sigma^{-1}(x-\mu_1, y-\mu_2)^t = \\ =& \frac{1}{|\Sigma|}(x-\mu_1, y-\mu_2)\begin{pmatrix} \sigma_y^2 & -\rho\sigma_x\sigma_y \\ -\rho\sigma_x\sigma_y & \sigma_x^2 \end{pmatrix}(x-\mu_1, y-\mu_2)^t = \\ =& \frac{1}{|\Sigma|}(x-\mu_1, y-\mu_2)\begin{pmatrix} (x-\mu_1)\sigma_y^2 - (y-\mu_2)\rho\sigma_x\sigma_y \\ (y-\mu_2)\sigma_x^2 - (x-\mu_1)\rho\sigma_x\sigma_y \end{pmatrix} = \\ =& \frac{(x-\mu_1)^2\sigma_y^2 + (y-\mu_2)^2\sigma_x^2 - 2(y-\mu_2)(x-\mu_1)\rho\sigma_x\sigma_y}{(1-\rho^2)\sigma_x^2\sigma_y^2} \end{aligned}$$

Note now that

$$Q(x,y) - \frac{(x-\mu_1)^2}{\sigma_x^2} = \frac{(x-\mu_1)^2\rho^2\sigma_y^2 + (y-\mu_2)^2\sigma_x^2 - 2(y-\mu_1)(x-\mu_1)\rho\sigma_x\sigma_y}{(1-\rho^2)\sigma_x^2\sigma_y^2} = \frac{\big(y-\mu_2-(x-\mu_1)\frac{\sigma_y}{\sigma_x}\rho\big)^2}{(1-\rho^2)\sigma_y^2}$$

Hence, the probability density function of Y given that $X = x$ is, using the previous Lemma

$$\begin{aligned} f_{(Y|X=x)}(y) =& \frac{f_{(X,Y)}(x,y)}{f_X(x)} = \frac{\frac{1}{2\pi|\Sigma|^{1/2}}\exp\{-\frac{Q(x,y)}{2}\}}{\frac{1}{\sqrt{2\pi}\sigma_x}\exp\{-\frac{(x-\mu_1)^2}{2\sigma_x^2}\}} = \frac{1}{\sqrt{2\pi(1-\rho^2)\sigma_y^2}}\exp\{-\frac{1}{2}\big(Q(x,y) - \frac{(x-\mu_1)^2}{\sigma_x^2}\big)\} = \\ =& \frac{1}{\sqrt{2\pi(1-\rho^2)\sigma_y^2}}\exp\{-\frac{\big(y-\mu_2-(x-\mu_1)\frac{\sigma_y}{\sigma_x}\rho\big)^2}{2(1-\rho^2)\sigma_y^2}\} \end{aligned}$$

which is the probability density function of a $N\Big(\mu_2 + \frac{\sigma_y}{\sigma_x}\rho(x-\mu_1), (1-\rho^2)\sigma_y^2\Big)$, as needed. $\square$

Remark. *Note that, given Theorem 3, $E[(Y|X=x)] = \mu_2 + \frac{\sigma_y}{\sigma_x}\rho(x-\mu_1) = \alpha + \beta x$, with $\alpha = \mu_2 - \mu_1\rho\frac{\sigma_y}{\sigma_x}$ and $\beta = \rho\frac{\sigma_y}{\sigma_x}$. This, in turn, can be rewritten as*

$$\alpha = \mu_Y - \mu_X\frac{\sigma_{xy}}{\sigma_x^2} \qquad \textit{and} \qquad \beta = \frac{\sigma_{xy}}{\sigma_x^2}$$

2 Particular case study

Let us consider now a particular case in which X is the random variable associated to an individual's height in cm and Y represents the random variable of the individual's weight in kg.

2.1 Sample generation

Assume the random vector (X, Y) follows a bidimensional normal distribution $N_2(\vec{\mu}, \Sigma)$ where $\vec{\mu}$ and Σ are such that

$$\vec{\mu} = \begin{pmatrix} 170 \\ 62 \end{pmatrix} \qquad \Sigma = \begin{pmatrix} 64 & 54 \\ 54 & 121 \end{pmatrix}$$

Given $v = \begin{pmatrix} a \\ b \end{pmatrix} \in \mathbb{R}^2$ one can compute

$$v^T \Sigma v = (a, b) \begin{pmatrix} 64 & 54 \\ 54 & 121 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = 64a^2 + 108ab + 121b^2 = 28a^2 + (6a + 9b)^2 + 40b^2 > 0 \iff (a, b) \neq (0, 0)$$

from which one concludes Σ is a positive definite matrix over $M_2(\mathbb{R})$. Hence, by [3], one can straightforwardly see that Σ can be factored as $\Sigma = AA^T$ where $A \in M_2(\mathbb{R})$ is a lower triangular matrix with positive diagonal elements. This is called the Cholesky decomposition of Σ and it is unique. In order to calculate A, the algorithm proposed in [3] has been followed. Using the notation defined in 2 for A,

$$a_{11} = \sqrt{64} = 8, \quad a_{12} = 0, \quad a_{21} = \frac{54}{a_{11}} = \frac{27}{4}, \quad a_{22} = \sqrt{121 - a_{12}^2} = \sqrt{\frac{1207}{16}} \quad \Longrightarrow A = \begin{pmatrix} 8 & 0 \\ 27/4 & \sqrt{1207/16} \end{pmatrix}$$

Hence, clearly

$$X = 170 + 8 \cdot Z_1 \qquad Y = 62 + \frac{27}{4} \cdot Z_1 + \sqrt{\frac{1207}{16}} \cdot Z_2$$

with $Z_1 \approx N(0, 1)$ and $Z_2 \approx N(0, 1)$. By generating random samples of Z_1 and Z_2, random samples for both X and Y are created. Let us denote the obtained sample as a set of points in the plane $\{(x_i, y_i)\}_{i=1}^n$ where n is the size of the sample.

Given what has been previously showed, the intercept α_0 and slope β_0 of the line $\mu_{Y|X} = E[(Y|X = x)] = \alpha_0 + \beta_0 x$ are:

$$\alpha_0 = \mu_Y - \frac{\sigma_{xy}}{\sigma_x^2} \mu_X = -\frac{1303}{16} \qquad \beta_0 = \frac{\sigma_{xy}}{\sigma_x^2} = \frac{27}{32}$$

Note that these coefficients have been easily obtained because the parameters $\vec{\mu}$ and Σ were known. Had they been unknown, α_0 and β_0 ought to have been estimated. In particular, one can use the method of least squares [5] or the maximum likelihood estimation [6] to estimate the coefficients α and β of regression line $y = \alpha + \beta x$ from the sample. Afterwards, one can show these estimators are compatible with the coefficients α_0 and β_0 predicted theoretically. This can be done by considering the following hypothesis testings.

$$\begin{cases} H_0 : \alpha = \alpha_0 \\ H_1 : \alpha \neq \alpha_0 \end{cases} \qquad \text{and} \qquad \begin{cases} H_0 : \beta = \beta_0 \\ H_1 : \beta \neq \beta_0 \end{cases}$$

For the sake of simplicity, it will be assumed that the tests are independent.

As it has been previously explained, there are plenty of methods to estimate the values of α and β. In this paper, the least squares estimators will be used, denoted by $\hat{\alpha}$ and $\hat{\beta}$ respectively. If one represents the sample mean for the x_i and y_i values as $\overline{x}$ and $\overline{y}$ respectively, one gets [4]

$$\hat{\beta} = \frac{S_{xy}}{S_x^2} \quad \text{and} \quad \hat{\alpha} = \overline{y} - \hat{\beta}\overline{x}$$

where $S_{xy} = \frac{1}{n} \sum_i^n (x_i - \overline{x})(y_i - \overline{y})$ and $S_x = \frac{1}{n} \sum_i^n (x_i - \overline{x})^2$.

Let us finally define $\text{SSE} = \sum_i^n (y_i - \hat{\alpha} - \hat{\beta}x_i)^2$ and

$$S^2 = \hat{\sigma}^2 = \frac{\text{SSE}}{n-2}$$

which will be later used.

2.2 Simulation results and analysis

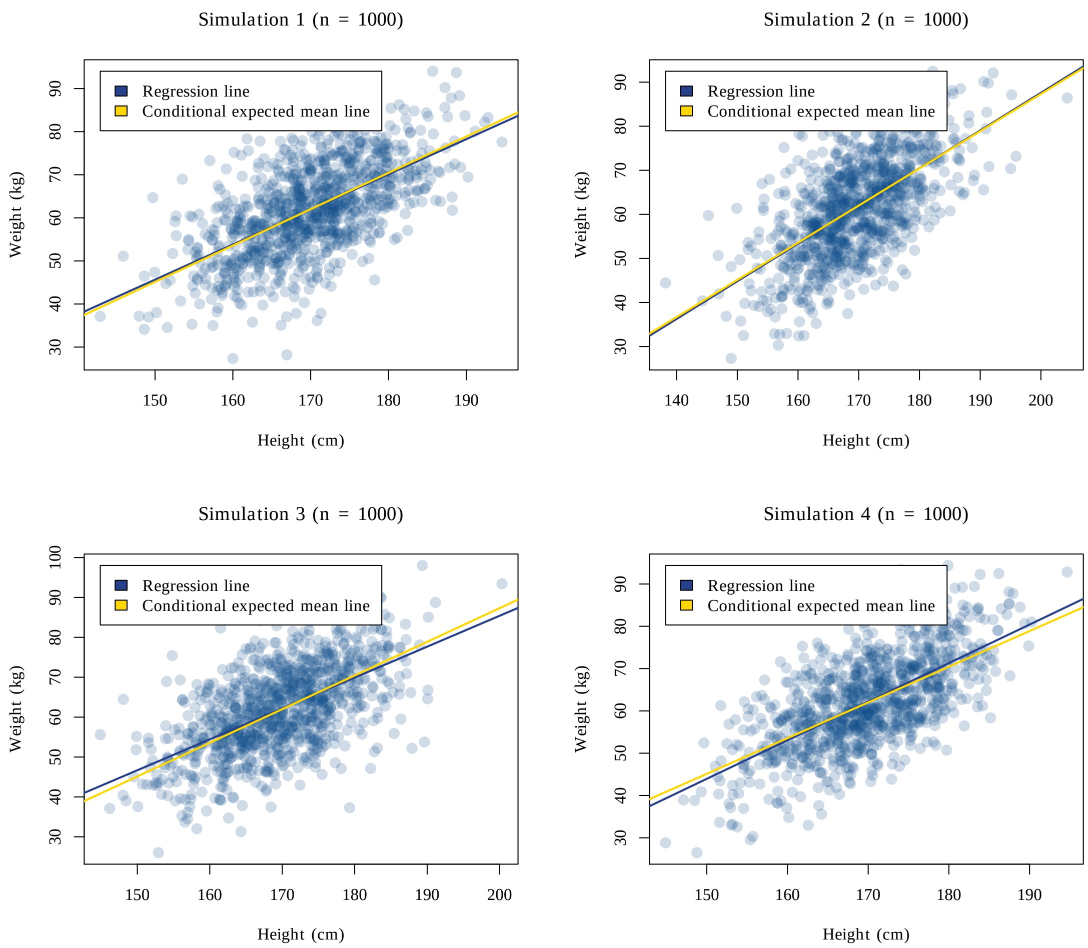

Figure 1: Simulations run in `R code` with sample size $n = 1000$.

Figure 1 shows four representative simulations. As expected, the empirical calculation of the regression line is in most cases a reliable estimation of the theoretical line. In order to proof the estimations are compatible with the theoretically predicted values, the above described tests have been run. It is known that [4]

$$\frac{\hat{\alpha} - \alpha}{\text{se}(\hat{\alpha})} \approx t_{n-2} \quad \text{and} \quad \frac{\hat{\beta} - \beta}{\text{se}(\hat{\beta})} \approx t_{n-2}$$

where t_{n-2} is a random variable following a $t-$Student distribution with $n-2$ degrees of freedom. The expressions $\text{se}(\hat{\alpha})$ and $\text{se}(\hat{\beta})$ represent the estimated errors, which are defined as

$$\text{se}(\hat{\alpha}) = \hat{\sigma}\sqrt{\frac{1}{n}\left(1 + \frac{\overline{x}^2}{S_x^2}\right)} \quad \text{and} \quad \text{se}(\hat{\beta}) = \frac{\hat{\sigma}}{S_x\sqrt{n}}$$

A significance level of $\gamma = 0.05$ has been set for our tests. For each test, the p-value for the following estimators has been computed

$$\frac{\hat{\alpha} - \alpha_0}{\mathrm{se}(\hat{\alpha})} \quad \text{and} \quad \frac{\hat{\beta} - \beta_0}{\mathrm{se}(\hat{\beta})}$$

Since our null hypotheses are $\alpha = \alpha_0$ and $\beta = \beta_0$, if the calculated p-values exceed the preset significance level, H_0 are accepted with a confidence level of 95%.

10^5 samples as the ones described have been drawn in order to estimate the ratio of simulations in which the defined tests fail (the test is said to fail when the p-value is lower than the significance level, resulting in H_0 being rejected). The results obtained indicate that in 5.11% of the cases H_0 has to be rejected for the α test. For β, the failing ratio is 5.07%. Given the similarities between the presented values, it is straightforward to ask oneself in how many simulations both tests fail simultaneously. It has been observed that in the 4.86% of the simulations run the null hypothesis had to be rejected in both tests simultaneously. This shows that the run tests are not fully independent, but can be studied separately.

This should not be an unexpected result. Indeed, note that the regression line passes through the point $(\overline{x}, \overline{y})$, which is expected to be pretty close[1] to the point (μ_X, μ_Y). This last point passes through the line $\mu_{Y|X} = \alpha_0 + \beta_0 x$. Then, any big change in the slope will cause the intercept to move along the axis and vice-versa. This effect is less noticeable as $(\overline{x}, \overline{y}) \to (0, 0)$, point in which even if the slope changes, the intercept remains in place.

Note that the significance level of both hypothesis tests has been chosen to be $\alpha = 5\%$. Hence, one would expect about 5% of false positives, i.e. about 5% of all simulations in which the null hypothesis is rejected although it is actually true. Hence, since the ratios of failed tests are around 5%, it can be argued that most of these could actually be false positives.

2.3 Confidence and prediction intervals

In previous projects (see [5] and [6]) it has already been studied how to calculate both confidence and prediction intervals with a certain level of confidence (see [4] for further reading). As in previous paragraphs, let us set the significance level to be $\gamma = 0.05$.

The $1 - \gamma$ confidence interval of $\mu_{Y|X}$ when X takes the value x_0 is

$$\left(\hat{\alpha} + \hat{\beta} x_0 - t_{n-2,1-\frac{\gamma}{2}} \cdot \hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{nS_x^2}}, \hat{\alpha} + \hat{\beta} x_0 + t_{n-2,1-\frac{\gamma}{2}} \cdot \hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{nS_x^2}}\right)$$

Thus, taking a particular simulation where H_0 is accepted with a 95% yielding $\hat{\alpha} = -87.66936$ and $\hat{\beta} = 0.8792187$, the confidence interval for Y knowing that $X = x_0 = 168$ cm is

$$(59.47855, 60.60022)$$

The prediction interval in the same scenario is broader and it reads

$$\left(\hat{\alpha} + \hat{\beta} x_0 - t_{n-2,1-\frac{\gamma}{2}} \cdot \hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{nS_x^2}}, \hat{\alpha} + \hat{\beta} x_0 + t_{n-2,1-\frac{\gamma}{2}} \cdot \hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{nS_x^2}}\right)$$

For the same particular estimator values, with a 95% confidence, the weight of a 168cm tall individual will lie in

$$(42.72474, 77.35403)$$

[1]By pretty close we mean that, as a matter of fact, $E(\overline{x}) = \mu_X$ and $E(\overline{y}) = \mu_Y$.

3 Conclusions

Let (X, Y) be a bidemensional normally distributed random vector defined by

$$\vec{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \qquad \Sigma = \begin{pmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{pmatrix}$$

Then, it has been shown that the conditional distribution of Y given that $X = x$, $(Y|X = x)$, follows a normal distribution, with expected value $\mu_2 + \frac{\sigma_{xy}}{\sigma_x^2}(x - \mu_1)$. Hence, one can find a linear regression to predict the value of Y given that $X = x$.

The theoretical discussion has been used in a particular case, relating the height X of an individual and its weight Y. Given $\vec{\mu}$ and Σ, samples of $n = 10^3$ vectors have been drawn and, by the method of least squares, the parameters α and β giving the regression line have been estimated by $\hat{\alpha}$ and $\hat{\beta}$. Independent hypothesis testings have been run to determine whether the hypothesis $\hat{\alpha} = \alpha_0 := \mu_2 - \mu_1 \frac{\sigma_{xy}}{\sigma_x^2}$ and $\hat{\beta} = \beta_0 := \frac{\sigma_{xy}}{\sigma_x^2}$ are compatible with our samples. With a significance level of 5%, it has been found that, for both tests, the null hypothesis has to be rejected in about 5.4% of the 10^5 drawn samples. In addition, in about 5.2% of drawn samples, both hypothesis ought to be rejected. This makes us believe that both tests are not really independent although they were treated as such.

Finally, a sample for which both null hypothesis were accepted has been picked. A confidence and a prediction intervals have been calculated for Y given a particular value of X. It has been noted than the prediction interval is wider that the confidence interval, as expected, since the error term variance σ^2 has to be added.

References

[1] J. del Castillo, G. Binotto, Universitat Autònoma de Barcelona. *Normals bidimensionals i regressió lineal*, Campus Virtual de la UAB (2020)

[2] M. Jolis, Universitat Autònoma de Barcelona. *Probabilitat i modelització estocàstica* (2020)

[3] L. Vandenberghe. University of California, Los Angeles. *Cholesky factorization* (2019) `http://www.seas.ucla.edu/ vandenbe/133A/lectures/chol.pdf`

[4] G. Binotto, Universitat Autònoma de Barcelona. *Pràctica 6: Regressió lineal simple*, Campus Virtual de la UAB (2020)

[5] A. Marín, A. Rodas, M. Seguí. *Estadística: Regressió lineal* (2020)

[6] A. Marín, A. Rodas, M. Seguí. *Estadística: Regressió lineal II* (2020)