Simple Regression

Anis Rezgui
Mathematics Department
Carthage University - INSAT
Tunis - Tunisia

March 28, 2022

Table of contents

The problem

Let x and y be two statistical variables:

$$x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \quad \text{and} \quad y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}.$$

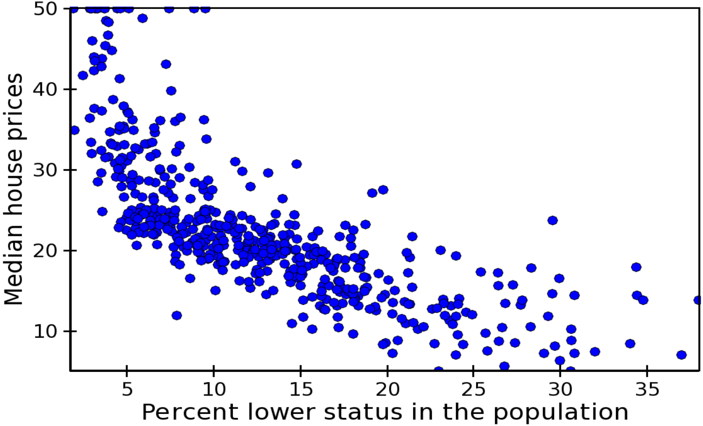
- The correlation problem, in general, consists of looking for a possible relationship between x and y: $y = f(x) + \text{"}\textit{residual}\text{"}$ that makes the residual part the smallest possible.
- The variable x is called "predictor" or the "independent" variable, y is called "response" or "dependent" variable.
- If we look for "linear" functional "f", the correlation is hence of linear type.

The scattergram

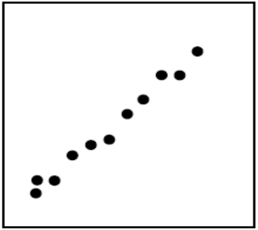
The sattergram or "scatter-plot" is simply the set of points of coordinates

$$\{(x_i, y_i) : i = 1, \cdots, n\}.$$

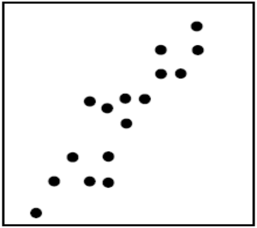
It gives a very useful a priori glance:



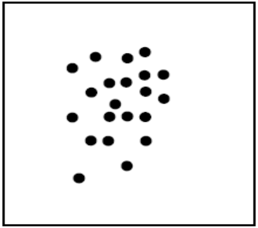
Different types of correlation



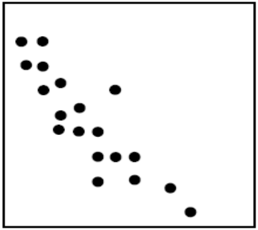
Strong positive correlation



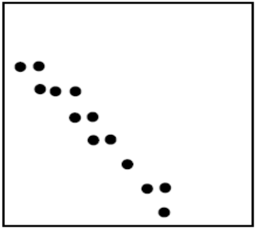
Moderate positive correlation



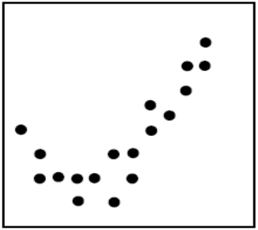
No correlation



Moderate negative correlation



Strong negative correlation



Curvilinear relationship

The least square distance

Suppose our problem reduced to the linear case, then it can be reformulated as follows:
We look for β and α such that if $y_i^* = \beta x_i + \alpha$ then we have

$$\overline{y^*} = \overline{y}$$
$$\sum_{i=1}^{n} (y_i^* - y_i)^2 \text{ is minimal.}$$

1. The line $s = \beta t + \alpha$ is called "least square line" or "regression line".
2. β is called "slope" and α the "intercept" of the regression line.

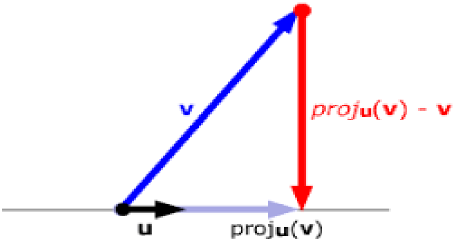
Vectorial formulation

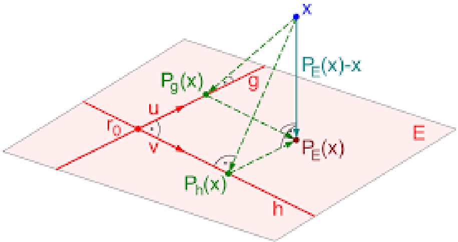
We look for a vector $y^* \in \mathbb{R}^n$ spanned by $\{x, \mathbf{1}\}$ i.e find out β and α such that $y^* = \beta x + \alpha \mathbf{1}$ minimizes the Euclidian norm $\|y - y^*\|_n^2$.

Theorem

$\|y - y^*\|_n^2$ *is minimal if and only if*
$y^* =$ *Orthogonal Projection of* y *on the subspace span*$\{x, \mathbf{1}\}$.

Least square method





ANOVA

Concequences

- $y - y^*$ *is orthogonal to* y^*.

$$y = y^* \oplus^{\perp} y - y^* \Rightarrow \|y\|_n^2 = \|y^*\|_n^2 + \|y - y^*\|_n^2 \qquad (1)$$

- *since* $\overline{y} = \overline{y^*}$ *and (1) we obtain*
  $\Rightarrow \|y - \overline{y}\|_n^2 = \|y^* - \overline{y}\|_n^2 + \|y - y^*\|_n^2$
- *Total sum of squares (tss) = fitted sum of squares (fss) + residual sum of squares (rss)*
- $\Rightarrow s_y^2 = s_{y^*}^2 + s_{y-y^*}^2$
  Total variance = Explained variance + Residual variance

Determination of the regression slope and intercept

We look for β and α so that $y - y^*$ is orthogonal to $span\{x, \mathbf{1}\}$ i.e

1. On one hand

$$\begin{aligned} y - y^* \perp x &\Rightarrow {}^t x(y - y^*) = 0 \\ &\Rightarrow {}^t xy - \beta {}^t xx + \alpha n\overline{x} = 0 \\ &\Rightarrow |x|^2\beta - n\overline{x}\alpha = {}^t xy \end{aligned} \tag{2}$$

2. on the other hand

$$\begin{aligned} y - y^* \perp \mathbf{1} &\Rightarrow {}^t\mathbf{1}(y - y^*) = 0 \\ &\Rightarrow n\overline{y} - n\beta\overline{x} + n\alpha = 0 \\ &\Rightarrow \alpha = \beta\overline{x} - \overline{y} \end{aligned} \tag{3}$$

Finally

Combining (2) and (3)

$$\begin{cases} \beta = \dfrac{\overline{x * y} - \overline{x} * \overline{y}}{\overline{x^2} - \overline{x}^2} = \dfrac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \\ \\ \alpha = \overline{y} - \beta\overline{x} \end{cases}$$

Linear Detrmination coefficient vs Correlation Coefficient

1. Set the Linear Determination Coefficient as

$$L^2_{y|x} = \frac{fss}{tss} = \frac{\sum_i (y_i^* - \overline{y})^2}{\sum_i (y_i - \overline{y})^2} \in [0,1]$$

2. It represents the strongness/strength of the linear correlation between y and x.

3. It can be read as the the rate of y explained linearly by x.

Linear Detrmination Coefficient vs Correlation Coefficient

Definition

The linear correlation coefficient of the statistical variables x and y is given by

$$
\begin{aligned}
r_{xy} &= \frac{\mathrm{Cov}(x,y)}{\sqrt{\sum_i (x_i-\overline{x})^2 \sum_i (y-\overline{y})^2}} && (4)\\
&= \frac{\sum_i (x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum_i (x_i-\overline{x})^2 \sum_i (y-\overline{y})^2}} && (5)\\
&= \frac{n}{n-1}\frac{\overline{x*y}-\overline{x}*\overline{y}}{s_x * s_y} && (6)\\
&= \frac{\sum_i x_i y_i - \frac{(\sum_i x_i)(\sum_i y_i)}{n}}{\sqrt{\left(\sum_i x_i^2 - \frac{(\sum_i x_i)^2}{n}\right)\left(\sum_i y_i^2 - \frac{(\sum_i y_i)^2}{n}\right)}} && (7)
\end{aligned}
$$

Very Important Remark

1. The linear correlation coefficient can be seen as the cosine of the angle formed by the two vectors of $\mathbb{R}^n$, $x - \overline{x} * \mathbf{1}$ and $y - \overline{y} * \mathbf{1}$:

$$\begin{aligned} \cos\left(x - \overline{x} * \widehat{\mathbf{1}, y} - \overline{y} * \mathbf{1}\right) &= \frac{{}^t(x - \overline{x} * \mathbf{1})(y - \overline{y} * \mathbf{1})}{\|x - \overline{x} * \mathbf{1}\| \|y - \overline{y} * \mathbf{1}\|} \\ &= \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}} \end{aligned}$$

2. why?

$$r_{xy}^2 = L_{y|x}^2. \tag{8}$$

Very Important Remark

1. The linear correlation coefficient can be seen as the cosine of the angle formed by the two vectors of $\mathbb{R}^n$, $x - \overline{x} * \mathbf{1}$ and $y - \overline{y} * \mathbf{1}$:

$$\cos\left(x - \overline{x} * \widehat{\mathbf{1}, y} - \overline{y} * \mathbf{1}\right) = \frac{{}^t(x - \overline{x} * \mathbf{1})(y - \overline{y} * \mathbf{1})}{\|x - \overline{x} * \mathbf{1}\| \|y - \overline{y} * \mathbf{1}\|} = \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^n (x_i - \overline{x})^2 \sum_{i=1}^n (y_i - \overline{y})^2}}$$

2. why?

$$r_{xy}^2 = L_{y|x}^2. \tag{8}$$

   1. We have looked only for the best linear function of "x" that describes "y", it is not possible to look for all possible functions of "x", since we only have samples.

Consequences

1. Since the correlation coefficient is a cosine it satisfies:

$$|r_{xy} = \sqrt{L_{xy}^2}| \leq 1.$$

2. If $|r_{xy}| = 1$ it means that we have a perfect linear correlation:

$$y_i = \beta x_i + \alpha, \quad \text{for all } i = 1, \cdots, n.$$

3. If $r_{xy} = 0$ it means that there is no linear relationship between x and y.
4. In all cases we have

$$\beta = r_{xy} \times \sqrt{\frac{\sum_{i=1}^{n}(y_i - \overline{y})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}}$$

We already know

1. If $|r_{xy}| \in [0, 25\%[$, there is no linear correlation !
2. If $|r_{xy}| \in [25\%, 50\%[$, there is a moderate linear correlation.
3. If $|r_{xy}| \in [50\%, 75\%[$, there is a fair linear correlation.
4. If $|r_{xy}| \in [75\%, 100\%[$, there is good linear correlation.

Inferential study

Suppose that the two statistical series x and y came from two given random variables X and Y and let $\{X_i \ : \ i = 1 \cdots n\}$ and $\{Y_i \ : \ i = 1 \cdots n\}$ be two random samples of X and Y. Denote by

$$TSS = \sum_{i=1}^{n}(Y_i - \overline{Y})^2, \quad FSS = \sum_{i=1}^{n}(Y_i^* - \overline{Y})^2$$

and

$$RSS = \sum_{i=1}^{n}(Y_i - Y_i^*)^2$$

and consider the following statistic:

$$F = \frac{FSS}{RSS}(n-2).$$

A theoretical result: The linear Gaussian model

Theorem

Suppose that X and Y are normally distributed and that they are independent. Then

$$\frac{FSS}{RSS}(n-2) = \mathcal{F}(1, n-2)$$

where $\mathcal{F}(1, n-2)$ is the Fisher distribution of degrees of freedom 1 and $n-2$.

Proposition and Definition

If $A \sim \chi^2_{d_1}$ and $B \sim \chi^2_{d_2}$ and are independent then

$$\frac{A/d_1}{B/d_2} \sim \mathcal{F}(d_1, d_2).$$

So what: How to use the latter Theorem ?

We consider the test: (H_0): $\beta = 0$ (which means that there is no linear relationship between X and Y) against (H_1): $\beta \neq 0$:

- we reject the hypothesis (H_0) and accept (H_1) with a confidence level of 95%, if:

$$f^* = \frac{fss}{rss}(n-2)$$

$$\mathbb{P}\{\mathcal{F}(1, n-2) \leq f^*\} = 95\%$$

- or equivalently (and in general) we reject H_0 and accept H_1 when

$$p - value = \mathbb{P}\{\mathcal{F}(1, n-2) \geq f^*\} \ll 1$$

Confidence interval for the regression line

Let $T_{n-2,\gamma/2}$ be such that $\mathbb{P}\{|\mathcal{T}_{n-2}| > T_{n-2,\gamma/2}\} = \gamma/2$.

1. A confidence interval of level $1-\gamma$, for the slope B is
$$\beta \pm T_{n-2,\gamma/2} \times \frac{s_y}{\sqrt{ssx}}.$$
2. A confidence interval of level $1-\gamma$, for the intercept A is
$$\alpha \pm T_{n-2,\gamma/2} \times s_y\sqrt{\frac{1}{n} + \frac{\overline{x}^2}{ssx}}.$$
3. The prediction interval for a new observation x_0, of level $1-\gamma$, is
$$\alpha + \beta x_0 \pm T_{n-2,\gamma/2} \times s_y\sqrt{1 + \frac{1}{n} + \frac{(x_0-\overline{x})^2}{ssx}}.$$

Student Distribution

Definition

Let T be a continuous random variable, it follows a t-student distribution with n degree of freedom if its density is given by

$$f(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\Gamma\left(\frac{n}{2}\right)}\left(1+\frac{t^2}{n}\right)^{-\frac{n+1}{2}}.$$

We denote $T \sim \mathcal{T}_n$.

Proposition

Let $Z \sim N(0,1)$ and $D \sim \chi^2_n$, suppose more that Z and D are independent. Then

$$T = \frac{Z}{\sqrt{D/n}} \sim \mathcal{T}_n.$$

Checking Gaussian hypothesis

To valid our results we need to check our Gaussian assumption about the residual:

1. We need to test if the residue comes from a normal distribution, for this we may use the QQ-plot graphical test.

2. We need to test if the residue and the fitted come from independent variables, for this we may use the scatterplot of the residue versus the fitted values as a graphical test.

Non-Gaussian case

- All results above are based on the assumption of normality of the residue.
- One may ask whether we still have the same results if the normality assumption is not any more satisfied?
- This leads to the mathematical investigation of looking at the limit behavior of the distribution when the number of observations goes to infinity.
- The results are still approximately correct !
- What we should do when the model is non-gaussian? A suggestion is:

  1. add predictor variable(s).
  2. and look for non-linear model.

- This will be the topic of the next chapters.