

*some title*

*January 5, 2018*

## *Regression*

### *Correlation*

The most crucial part in defining correlation geometrically is defining the dot product as it enables to compute the length of a vector:

$$|\vec{a}| = \sqrt{\langle \vec{a}, \vec{a} \rangle}$$

and the angle between any two vectors:

$$\cos(\vec{a}, \vec{b}) = \frac{\langle \vec{a}, \vec{b} \rangle}{|\vec{a}| |\vec{b}|}$$

Now we define scalar product of two random vectors as covariation between them:

$$\langle X, Y \rangle = \text{Cov}(X, Y)$$

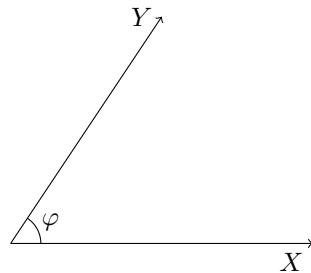
The main characteristics of a random vector are its length and direction. So, we introduce the length

$$\sqrt{\text{Cov } X, X} = \sqrt{\text{Var}(X)} = \sigma_X$$

and the angle between two random vectors

$$\cos(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X) \text{Var}(Y)}} = \text{Corr}(X, Y)$$

Note that from the definition of the angle it follows that correlation can range from  $-1$  to  $1$ .



The most widespread definition of correlation is

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X) \text{Var}(Y)}}$$

Figure 1: Random vector  $X$  of length  $\sigma_X$  and random vector  $Y$  of length  $\sigma_Y$ ,  $\cos \varphi$  is the angle between  $X$  and  $Y$ .

Another important geometrical tool is projection. Recall that for any two vectors the scalar product  $\langle \vec{a}, \vec{b} \rangle$  can be interpreted as the length of projected  $\vec{b}$  multiplied by the length of  $\vec{a}$ . The projection itself is  $\cos(\vec{a}, \vec{b})\vec{b}$ . Same holds for random vectors. The projection of a random vector  $Y$  onto  $\{cX | c \in \mathbb{R}\}$  is  $\hat{Y} = \text{Corr}(X, Y) \cdot Y$ .

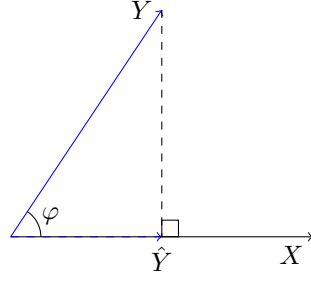


Figure 2: Random vector  $Y$  projected onto random vector  $X$ .

Looking at Figure , we can interpret the square of correlation coefficient. Using the fact that  $\cos^2 \varphi$  is the squared ratio of the leg adjacent to  $\varphi$  to hypotenuse, we can write

$$\text{Corr}^2(X, Y) = \frac{\text{Var}(\hat{Y})}{\text{Var}(Y)}$$

as the variance of a random vecor is associated with the square of its length. Thus, the correlation coefficient squared shows the fraction of variance in  $Y$  which can be explained with the most similar random variable proportional to  $X$ .

### Sample correlation coefficient in simple linear regression

**Theorem 1.** *A linear regression model with one explanatory variable and constant term has the property*

$$\text{sCorr}(y, \hat{y}) = \text{sign}(\hat{\beta}_2) \text{sCorr}(y, x)$$

*Proof.* Firstly, we consider the case when  $\hat{\beta}_2 > 0$  so the main picuture is of the form depicted in Figure . It has been shown earlier that the correlation coefficient represents the angle between two random vectors. So in order to complete the proof we need to find the appropriate angles and compare them.

However, it seems to be difficult to compare the angles in the three dimensional space. That is why we start with projecting both  $x$  and  $y$  onto the space perpendicular to the vector of all ones  $\mathbf{1}$  as shown in Figure 4(a). We denote this space as  $\text{Lin}^\perp(\mathbf{1})$ . The resulting vectors are  $x - \bar{x} \cdot \mathbf{1}$  and  $y - \bar{y} \cdot \mathbf{1}$  respectively since projection of any vector  $\vec{a}$  on the line given by a vector of all ones yields the vector of averages  $\vec{a}$ .

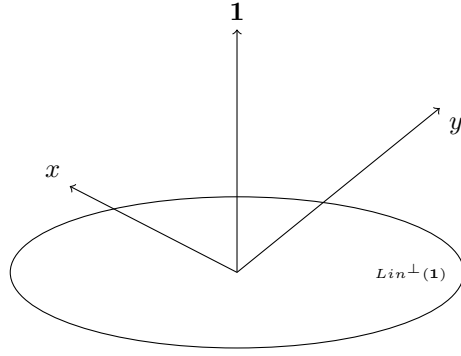
In order to get the angle between  $y$  and  $\hat{y}$  we should start with regressing  $y$  on  $\text{Lin}(x, \mathbf{1})$ . Then the only thing thing left is to project  $\hat{y}$  onto  $\text{Lin}^\perp \mathbf{1}$  since the  $y$  vector has already been projected. Note that the projected  $\hat{y}$  falls onto tha span of vector  $x - \bar{x} \cdot \mathbf{1}$  as it can be decomposed into a sum  $ax + b\mathbf{1}$  where  $a, b \in \mathbb{R}$ .  $ax$  is projected in the same way as  $x$  and  $b\mathbf{1}$  yields zero

Assuming the underlying relationship between  $x$  and  $y$  to be

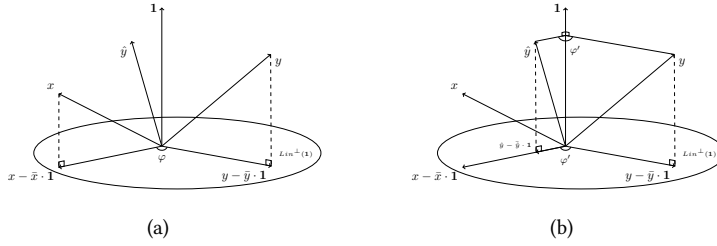
$$y_i = \beta_1 + \beta_2 x_i + \varepsilon_i \quad i = 1, \dots, n$$

where  $\varepsilon_i$  is an error term the following holds

$$\begin{aligned} \text{sCorr}(y, \hat{y}) &= \frac{\text{sCov}(y) \text{sCov}(\hat{y})}{\sqrt{\text{sVar}(y) \text{sVar}(\hat{y})}} \\ &= \frac{\text{sCov}(y) \text{sCov}(\hat{\beta}_1 + \hat{\beta}_2 x)}{\sqrt{\text{sVar}(y) \text{sVar}(\hat{\beta}_1 + \hat{\beta}_2 x)}} \\ &= \frac{\text{sCov}(y) \text{sCov}(\hat{\beta}_2 x)}{\sqrt{\text{sVar}(y) \text{sVar}(\hat{\beta}_2 x)}} \\ &= \frac{\hat{\beta}_2 \text{sCov}(y) \text{sCov}(x)}{|\hat{\beta}_2| \sqrt{\text{sVar}(y) \text{sVar}(x)}} \\ &= \text{sign}(\hat{\beta}_2) \frac{\text{sCov}(y) \text{sCov}(x)}{\sqrt{\text{sVar}(y) \text{sVar}(x)}} \end{aligned}$$

Figure 3: Vectors  $x$ ,  $y$  and  $1$ .

when projected onto the orthogonal space. The result of this step is shown in Figure 4(b).

Figure 4: (a): 'Centred'  $x$  and  $y$ , i.e., projected onto  $Lin^\perp(1)$ ; (b): 'Centred'  $\hat{y}$ , i.e., projected onto  $Lin^\perp(1)$ .

Since the projection of  $\hat{y}$  lies exactly on the span of vector  $x - \bar{x} \cdot 1$ , we can conclude that  $\cos \varphi = \cos \varphi'$  and to put it another way  $\text{sCorr}(x, y) = \text{sCorr}(y, \hat{y})$ .

Now consider the case when  $\hat{\beta}_2 < 0$ . Note that the sign of  $\beta_1$  does not influence the correlation coefficient sign. The only difference is that now  $\hat{y}$  is projected onto the span of  $x - \bar{x} \cdot 1$  not on this vector itself while the projections of  $x$  and  $y$  remain the same. Looking at Figure we deduce that the angle between  $y$  and  $\hat{y}$  is complemet to the angle between  $x$  and  $y$ . Using trigonometric properties, we simplify  $\cos(180^\circ - \varphi) = -\cos \varphi$  which in turn implies  $\text{sCorr}(x, y) = -\text{sCorr}(y, \hat{y})$ .  $\square$

### Regression line and point of averages

**Theorem 2.** *The point of averages lies on the estimated regression line.*

*Proof.* For the geometrical proof it suffices to show that  $\hat{y}$  is a linear combination of the regressors, which is true by construction, and that  $\frac{1}{n} \sum_{i=1}^n \hat{y}_i = \frac{1}{n} \sum_{i=1}^n y_i$ . In order for the pictures to be more clear the proof will be presented for the case of two regressors.

The first step is regressing  $y$  on  $Lin(1, x)$ . As shown in Figure 6(a), we obtain  $\hat{y}$  as a linear combination of  $1$  and  $x$ . The next step is to regress both

If the regression contains the intercept, the following equation holds:

$$\begin{aligned} \hat{y} &= X\hat{\beta} = X(X^T X)^{-1} X^T y \\ &= X(X^T X)^{-1} X^T X\beta + X(X^T X)^{-1} X^T \varepsilon \end{aligned}$$

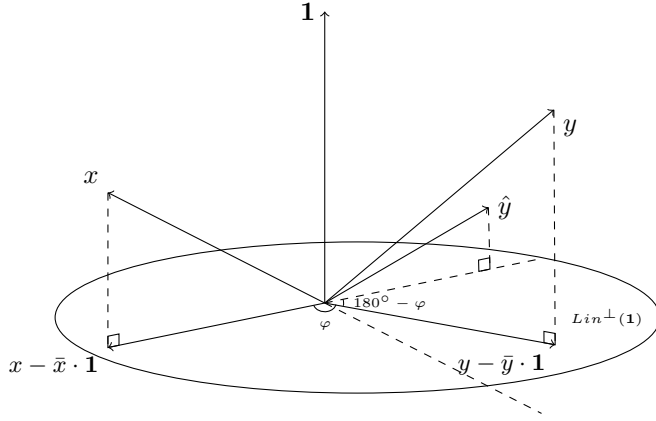
Premultiplying both sides by  $X^T$ , we obtain:

$$\begin{aligned} X^T \hat{y} &= X^T X(X^T X)^{-1} X^T X\beta \\ &\quad + X^T X(X^T X)^{-1} X^T \varepsilon \\ &= X^T X\beta + X^T \varepsilon \end{aligned}$$

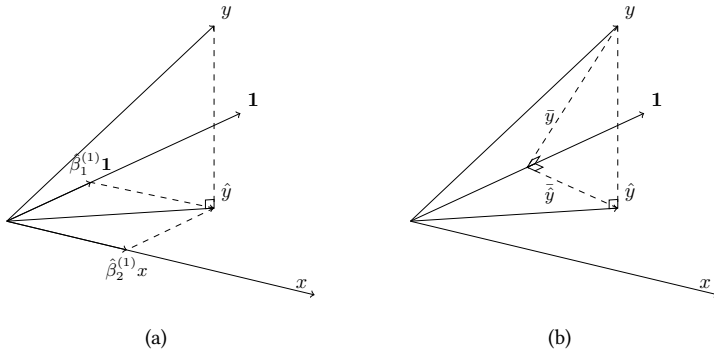
This is a system of equations. The first row of  $X^T$  is  $1$  vector, so we can write out the first equation:

$$\sum_{i=1}^n \hat{y}_i = \sum_{i=1}^n \sum_{j=1}^k x_{ij} \beta_j$$

From the first equation in the system

Figure 5: Case of  $\beta_2 < 0$ .

$y$  and  $\hat{y}$  on  $\mathbf{1}$  which results in  $\bar{y}$  and  $\hat{\bar{y}}$  correspondingly. By the theorem of three perpendiculars,  $\bar{y} = \hat{\bar{y}}$  which is shown in Figure 6(b).

Figure 6: (a): Regression of  $y$  on  $\text{Lin}(\mathcal{K}, x)$ ; (b): Regression of  $y$  and  $\hat{y}$  on  $\mathcal{K}$ .

□

### Frisch–Waugh–Lovell theorem

**Theorem 3.** Consider regression

$$y = X_1\beta_1 + X_2\beta_2 + u \quad (1)$$

where  $X_{n \times k} = [X_1 X_2]$ , i.e.  $X_1$  consists of first  $k_1$  columns of  $X$  and  $X_2$  consists of remaining  $k_2$  columns of  $X$ ,  $\beta_1$  and  $\beta_2$  are comfortable, i.e.  $k_1 \times 1$  and  $k_2 \times 1$  vectors. Consider another regression

$$M_1 y = M_1 X_2 \beta_2 + M_1 u \quad (2)$$

where  $M_1 = I - P_1$  projects onto the orthogonal complement of the column space of  $X_1$  and  $P_1 = X_1(X_1^T X_1)^{-1} X_1^T$  is the projection onto the column space of  $X_1$ . Then the estimate of  $\beta_2$  from regression 1 will be the same as the estimate from regression 2.

From regression 2 we get the following estimator:

$$\begin{aligned} \hat{\beta}_2 &= ((M_1 X_2)^T M_1 X_2)^{-1} (M_1 X_2)^T M_1 y \\ &= (X_2^T M_1^T M_1 X_2)^{-1} X_2^T M_1^T M_1 y \\ &= (X_2^T M_1 X_2)^{-1} X_2^T M_1 y \end{aligned}$$

As for regression 1, let us note that due to  $y = \hat{y} + \hat{u}$   $y$  can be decomposed as follows:

$$y = P y + M y = X_1 \hat{\beta}_1 + X_2 \hat{\beta}_2 + M y$$

Premultiplying both sides by  $X_2^T M_1$ , we obtain:

$$\begin{aligned} X_2^T M_1 y &= X_2^T M_1 X_1 \hat{\beta}_1 + X_2^T M_1 X_2 \hat{\beta}_2 + X_2^T M_1 M y \\ &= X_2^T M_1 X_2 \hat{\beta}_2 + X_2^T M_1 M y \\ &= X_2^T M_1 X_2 \hat{\beta}_2 \end{aligned}$$

On the last step we used the fact that

$$\begin{aligned} (X_2^T M_1 M y)^T &= y^T M^T M_1^T X_2 \\ &= y^T M M_1 X_2 = y^T M X_2 = 0^T \end{aligned}$$

Assuming  $X_2^T M_1 X_2$  is invertible, we get the same estimator

$$\hat{\beta}_2 = (X_2^T M_1 X_2)^{-1} X_2^T M_1 y$$

*Proof.* Geometrical proof will be presented for the following model:

$$y_i = \beta_1 x_i + \beta_2 z_i + u_i \quad (3)$$

We start with regression ‘all-at-once’ and will distinct its coefficients with index (1). The only step in obtaining  $\beta_1^{(1)}$  is regressing  $y$  on  $Lin(x, z)$  and then expanding  $\hat{y}$  as a linear combination of basis vectors  $x$  and  $z$ , which is shown in Figure 7(a). Figure 7(b) depicts  $Lin(x, z)$ .

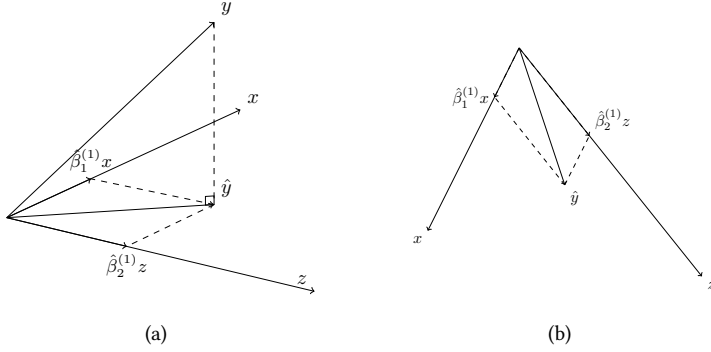


Figure 7: (a): Regression of  $y$  on  $Lin(x, z)$ ; (b):  $Lin(x, z)$ .

As for the model 2, where several regressions are performed consecutively, we start with regressing  $y$  on  $z$ , resulting in  $\tilde{y}$ , which we will refer to as ‘cleansed’  $y$ .

$$\begin{aligned} y &= \alpha z + \varepsilon \\ \hat{\alpha} &= \frac{y^T z}{z^T z} \\ \tilde{y} = \hat{\varepsilon} &= y - \frac{y^T z}{z^T z} z \end{aligned} \quad (4)$$

Following that,  $x$  is regressed on  $z$ , resulting in  $\tilde{x}$  – ‘cleansed’  $x$ .

$$\begin{aligned} x &= \gamma z + \nu \\ \hat{\gamma} &= \frac{x^T z}{z^T z} \\ \tilde{x} = \hat{\nu} &= x - \frac{x^T z}{z^T z} z \end{aligned} \quad (5)$$

Geometric results of these two steps are presented in 8(a).

Finally, ‘cleansed’  $y$  must be regressed on ‘cleansed’  $x$ . However, it cannot be performed immediately as  $\tilde{y}$  and  $\tilde{x}$  are skew lines. So at first, we fix this problem by translation and after that obtain  $\hat{\beta}_1^{(2)}\tilde{x}$  (see Figure 8(b)).

Now, let us picture all the results in one figure and mark some main points.

In Figure 9(b) segments AF and BH = DG stand for  $\hat{\beta}_1^{(1)}x$  and  $\hat{\beta}_1^{(2)}\tilde{x}$  respectively, while segments AC and BC represent  $x$  and  $\tilde{x}$ . Having two

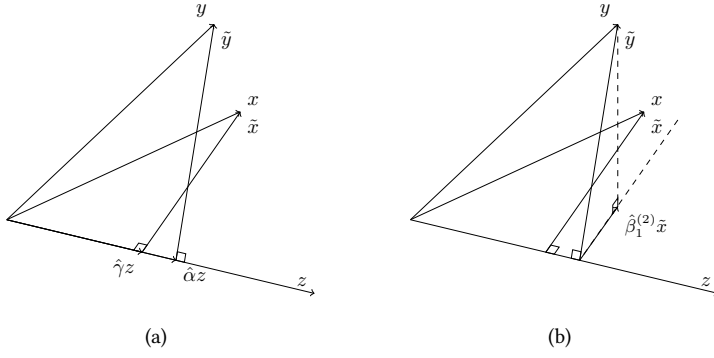


Figure 8: (a): Regression of  $y$  on  $z$  and of  $x$  on  $z$ ; (b): Translation of  $\tilde{x}$ .

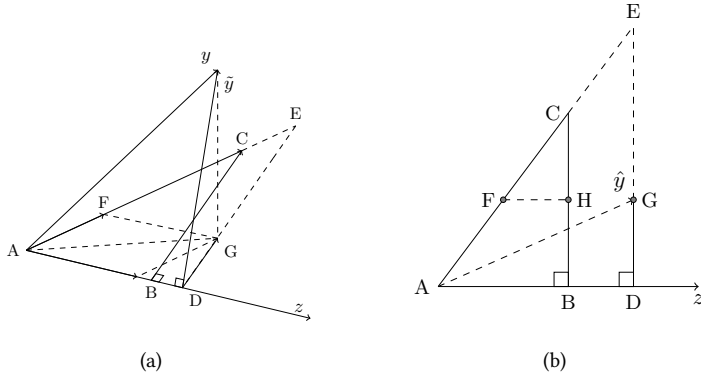


Figure 9: (a): Point A stands for the origin,  $B = \hat{\gamma}z$ ,  $C = x$ ,  $D = \hat{\alpha}z$ ,  $E$  — intersection of vector  $x$  and line parallel to  $\tilde{x}$ ,  $F = \hat{\beta}_1^{(1)}x$ ,  $G = \hat{\beta}_1^{(2)}\tilde{x}$ ; (b):  $Lin(x, z)$ .

congruent angles, triangles ABC and FHC are similar. Then, it follows:

$$\frac{AF}{AC} = \frac{BH}{BC} \Leftrightarrow \frac{\hat{\beta}_1^{(1)}x}{x} = \frac{\hat{\beta}_1^{(2)}\tilde{x}}{\tilde{x}} \Leftrightarrow \hat{\beta}_1^{(1)} = \hat{\beta}_1^{(2)}$$

□