Geometric interpretation of random variables

The fundamental part to start with is defining the geometric properties of random variables using some concepts of linear algebra.

Consider the vector space which consists of all random variables with finite mean and variance. We will regard each point in this space (or vector that correponds to that point in terms of linear algebra) a random variable. We define the scalar product of two random variables X and Y to be

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

It is of no difficulty to check that the definition satisfies the properties of scalar product assuming that X and Y are the same random variables if $\mathbb{P}(X = Y) = 1$.

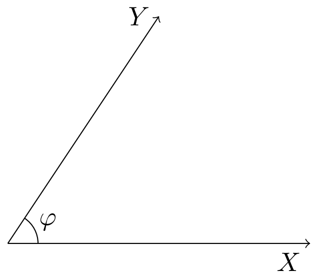


Figure 1: Geometric representation of random variables.

Having defined the inner product, we are now able to introduce the squared length of a random variable X which is

$$\|X\|^2 = \langle X, X\rangle = \operatorname{Cov}(X, X) = \operatorname{Var}(X),$$

so the length is simply the square root of this expression, i.e., the standard deviation of X (σ_X).

Recall that for any non-random vectors a and b the angle between them is calculated with the formula

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

The same applies for the random variables and it is already clear that two random variables are uncorrelated if and only if their scalar product equals to 0. Additionally, it means that these two random variables are orthogonal in the vector space.

The analogue for $cos(a, b)$ in the vector space of all the random variables is the correlation between two of them:

$$\operatorname{Corr}(X, Y) = \frac{\operatorname{Cov}(X, Y)}{\sqrt{\operatorname{Var}(X)\operatorname{Var}(Y)}} = \frac{\langle X, Y\rangle}{\sqrt{\|X\|^2\|Y\|^2}}.$$

From the equivalence of $\operatorname{Corr}(X, Y)$ to the $\cos(a, b)$ it automatically follows that the correlation coefficient can range from -1 to 1.

A useful property of the geoemtry of random variables is that all the geometric theorems still hold. For instance, the Pythogorean theorem can be formulated as follows: if the ranadom variables X and Y are uncorrelated (which implies that they are orthogonal), then the variance of their sum equals the sum of their variances:

$$\operatorname{Var}(X + Y) = \sigma^2_{X+Y} = \sigma^2_X + \sigma^2_Y = \operatorname{Var}(X) + \operatorname{Var}(Y).$$

Translated to the non-random language, assumption of uncorrelatedness correspnds to the right triangle setting, the variance of the sum of two

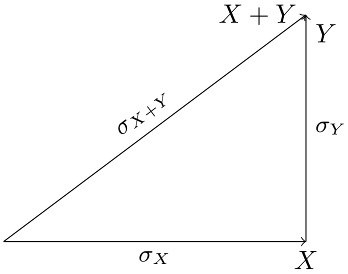


Figure 2: The Pythagorean theorem for random variables X and Y.

random variables stands for the hypotenuse squared and the sum of the variances is the sum of the legs squared.

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

Note that the squared lengths of the leg adjacent to φ and the hypotenuse are $\text{Var}(\hat{Y})$ and $\text{Var}(Y)$. So, the Figure 3 gives a useful expression for the correlation coefficient squared:

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

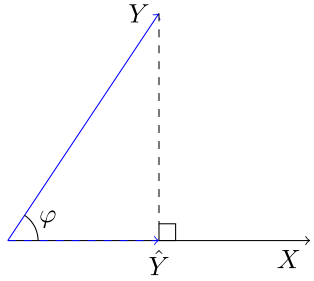


Figure 3: The projection of a random variable Y onto the line spanned by a random varibale X.

The law of iterated expectations

Theorem 1. *For any random variable X and Y,*

$$\text{E}(\text{E}(X|Y)) = \text{E}(Y).$$

Proof. Consider the vector space of all the random variables. The random variable which can be described as functions $h(X)$ of X form a subspace of that vector space, represented as a plane α in Figure . Another subspace is a subspace of constants, denoted as a vector $\mathbf{1} \in \alpha$.

In order to obtain $\text{E}(Y|X)$, first, we need to project Y onto the subspace corresponding to X. As a result of this step, we get $\text{E}(Y|X)$ — the function of X that predicts Y the best. Next, projecting $\text{E}(Y|X)$ onto the space of all constants, we end up with $E(Y)$.

Notice that the vector $Y - \text{E}(Y|X)$ (which is also called the residual) is perpendicular to the plane α. Moreover, the vecotr $\text{E}(Y|X) - E(Y)$ is perpendicular to the vector of constants $\mathbf{1}$. Thus, we can apply the theorem of three perpendiculars and conclude that the vector $Y - \text{E}(Y)$ is also perpendicular to the vector of constants $\mathbf{1}$.

So, we showed that the expectation of the random variable Y can be obtained either in two steps or by its direct projection onto the subpace of constans.

Here is the proof for the case when X and Y are both discrete. Let $\text{E}(Y|X) = g(X)$.

$$\begin{aligned}
\text{E}(g(X)) &= \sum_x g(x)\mathbb{P}(X = x) \\
&= \sum_x \left(\sum_y y\mathbb{P}(Y = y|X = x) \right) \mathbb{P}(X = x) \\
&= \sum_x \sum_y y\mathbb{P}(X = x)\mathbb{P}(Y = y|X = x) \\
&= \sum_y y \sum_x \mathbb{P}(X = x, Y = y) \\
&= \sum_y y\mathbb{P}(Y = y) \\
&= \text{E}(Y)
\end{aligned}$$

The proof in case of continous random variables is absolutely analogous.

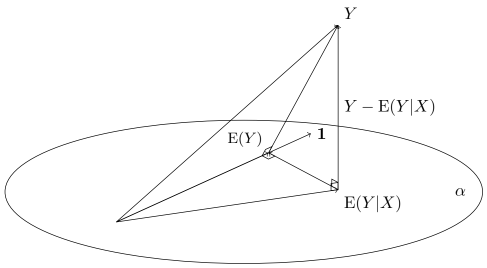


Figure 4: The law of iterated expectations. Equivalence of the two-step projecttion and direct projection of Y onto $\mathbf{1}$.

□

MSE decomposition

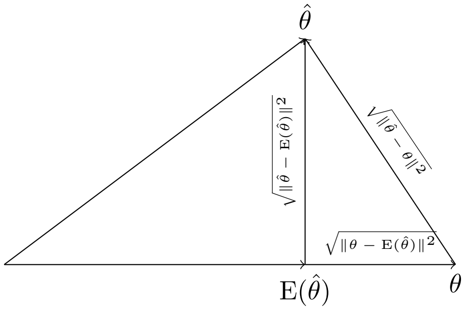
Theorem 2. *The mean squared error of an estimator $\hat{\theta}$ with respect to an unknown parameter θ defined as $MSE(\hat{\theta}) = \mathrm{E}((\hat{\theta} - \theta)^2)$ can be decomposed into the sum of the variance of the estimator and its squared bias:*

$$MSE(\hat{\theta}) = \mathrm{Var}(\hat{\theta}) + \mathrm{E}\left[\left(\mathrm{E}(\hat{\theta}) - \theta\right)^2\right]$$

Proof. We start with a random variable θ and its estimate $\hat{\theta}$ in the vector space. We know that an unbiased estimator's projection would be exactly the vector representing θ. However, in general it does not have to and Figure illustrates this case: the projection of the estimator falls onto the line spanned by the vector θ.

Connecting vectors θ and $\hat{\theta}$, we obtain the right triangle which legs are $\hat{\theta} - \mathrm{E}(\hat{\theta})$, $\mathrm{E}(\hat{\theta}) - \theta$ and the hypotenuse $\hat{\theta} - \theta$. Applying the Pythagorean theorem, we finish the proof:

$$\begin{aligned}
\|\hat{\theta} - \theta\|^2 &= \|\hat{\theta} - \mathrm{E}(\hat{\theta})\|^2 + \|\mathrm{E}(\hat{\theta}) - \theta\|^2 \\
\mathrm{E}((\hat{\theta} - \theta)^2) &= \mathrm{E}((\hat{\theta} - \mathrm{E}(\hat{\theta}))^2) + \mathrm{E}((\mathrm{E}(\hat{\theta}) - \theta)^2) \\
MSE(\hat{\theta}) &= \mathrm{Var}(\hat{\theta}) + \mathrm{E}((\mathrm{E}(\hat{\theta}) - \theta)^2)
\end{aligned}$$


$\hat{\theta}$
$\sqrt{\|\hat{\theta} - \mathrm{E}(\hat{\theta})\|^2}$
$\sqrt{\|\hat{\theta} - \theta\|^2}$
$\sqrt{\|\theta - \mathrm{E}(\hat{\theta})\|^2}$
$\mathrm{E}(\hat{\theta})$
θ

□

$$\begin{aligned}
MSE(\hat{\theta}) &= \mathrm{E}((\hat{\theta} - \theta)^2) \\
&= \mathrm{E}\left[\left(\hat{\theta} - \mathrm{E}(\hat{\theta}) + \mathrm{E}(\hat{\theta}) - \theta\right)\right] \\
&= \mathrm{E}\left[\left(\hat{\theta} - \mathrm{E}(\hat{\theta})\right)^2 + 2\left((\hat{\theta} - \mathrm{E}(\hat{\theta}))(\mathrm{E}(\hat{\theta}) - \theta)\right)\right. \\
&+ \left.\left(\mathrm{E}(\hat{\theta}) - \theta\right)^2\right] \\
&= \mathrm{E}\left[\left(\hat{\theta} - \mathrm{E}(\hat{\theta})\right)^2\right] \\
&+ 2\,\mathrm{E}\left[\left((\hat{\theta} - \mathrm{E}(\hat{\theta}))(\mathrm{E}(\hat{\theta}) - \theta)\right)\right] \\
&+ \mathrm{E}\left[\left(\mathrm{E}(\hat{\theta}) - \theta\right)^2\right] \\
&= \mathrm{E}\left[\left(\hat{\theta} - \mathrm{E}(\hat{\theta})\right)^2\right] \\
&+ 2(\mathrm{E}(\hat{\theta} - \mathrm{E}(\hat{\theta})))\,\mathrm{E}(\hat{\theta} - \mathrm{E}(\hat{\theta})) \\
&+ \mathrm{E}\left[\left(\mathrm{E}(\hat{\theta}) - \theta\right)^2\right] \\
&= \mathrm{E}\left[\left(\hat{\theta} - \mathrm{E}(\hat{\theta})\right)^2\right] + \mathrm{E}\left[\left(\mathrm{E}(\hat{\theta}) - \theta\right)^2\right] \\
&= \mathrm{Var}(\hat{\theta}) + \mathrm{E}\left[\left(\mathrm{E}(\hat{\theta}) - \theta\right)^2\right]
\end{aligned}$$

Figure 5: Decomposition of mean squred error into the variance and the bias squared.

Regression

The concepts discussed in the following section could also be presented in random variables instead of sample ones. As the geometry of sample variables is almost of no difference comparing to the random ones, the logic of all the theorems is also the same.

Geometry of sample variables

In the same manner which was performed in Section , we define the scalar product of two sample variables $x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$ and $y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$ as a sample covariation between them:

$$\langle x, y \rangle = \text{sCov}(x, y).$$

$$\text{sCorr}(x, y) = \frac{\text{sCov}(x, y)}{\sqrt{\text{sVar}(x)\,\text{sVar}(y)}}$$
$$= \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})\frac{1}{n-1}\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})}}$$

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

$$\sqrt{\text{sCov}(x, x)} = \sqrt{\text{sVar}(x)} = \sigma_x$$

and the angle between two sample variables

$$\cos(x, y) = \frac{\text{sCov}(x, y)}{\sqrt{\text{sVar}(x)\,\text{sVar}(y)}} = \text{sCorr}(x, y).$$

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

Completely analogus to the case of random variables, the projection of such a sample variable y onto $\{cx | c \in \mathbb{R}\}$ is $\hat{y} = \text{sCorr}(x, y) \cdot y$.

Looking at Figure 6, we can interpret the square of sample correlation coefficient. Using the fact that $cos^2\varphi$ is the squared ratio of the leg adjacent to φ to hypotenuse, we can conclude that

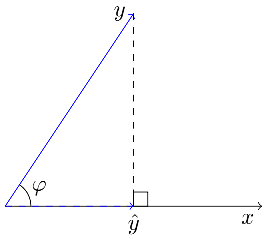


Figure 6: Vector y projected onto vector x.

$$\text{sCorr}^2(x, y) = \frac{\text{sVar}(\hat{y})}{\text{sVar}(y)},$$

as the variance of a vector is associated with the square of its length. Thus, the sample correlation coefficient squared shows the fraction of variance in y which can be explained with the most similar vector proportional to x.

Sample correlation when a constatnt vector added

Theorem 3. *Adding a vector of constants does not affect the sample correlation coefficient:*

$$\text{sCorr}(x + \alpha\mathbf{1}, y) = \text{sCorr}(x, y)$$

where $\alpha \in \mathbb{R}$.

$$\text{sCorr}(x + \alpha\mathbf{1}, y) = \frac{\text{sCov}(x + \alpha\mathbf{1}, y)}{\text{sVar}(x + \alpha\mathbf{1})\,\text{sVar}(y)}$$
$$= \frac{\text{sCov}(x, y)\,\text{sCov}(\alpha\mathbf{1}, y)}{\text{sVar}(x)\,\text{sVar}(y)}$$
$$= \frac{\text{sCov}(x, y)}{\text{sVar}(x)\,\text{sVar}(y)}$$
$$= \text{sCorr}(x, y)$$

Proof. Firstly, we project vectors x and y onto $Lin^{\perp}(\mathbf{1})$ in order to get $x^c = x - \bar{x}$ and $y^c = y - \bar{y}$ ('c' stands for 'centred'). It can be shown that the matrix corresponding to projecting onto the line spanned by a vector of all ones has the following form

$$\frac{\mathbf{1}^T\mathbf{1}}{\mathbf{1}\mathbf{1}^T} = \frac{\begin{pmatrix} 1 & \dots & 1 \end{pmatrix}\begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}}{\sum_{i=1}^{n} 1} = \begin{pmatrix} \frac{1}{n} & \dots & \frac{1}{n} \\ \vdots & \ddots & \vdots \\ \frac{1}{n} & \dots & \frac{1}{n} \end{pmatrix}$$

Thus, projecting onto the orthogonal subspace is equivalent to substracting the projected vector, i.e., the vector of averages, from the original one.

Also note that the angle φ between the original and centred vectors remains the same. The result of this step is shown in Figure 7.

Then we need to derive a new vector $\tilde{x}$ with constants added to each component. Geometrically adding a vector of costants means adding a vector of all ones scaled by $\alpha \in \mathbb{R}$, i.e., $\alpha\mathbf{1}$. Then the new vector $\tilde{x}$ can be broken up into a sum of $\alpha\mathbf{1}$ and βx, $\alpha, \beta \in \mathbb{R}$, which can be seen in Figure . After that we will project this new vector $\tilde{x}$ onto $Lin^{\perp}(\mathbf{1})$. By the properties of projection it is of no difference whether to project the whole vector $\tilde{x}$ or project its parts $\alpha\mathbf{1}$ and βx — the result is the same. So, while βx is projected onto the span of x^c, the projection of $\alpha\mathbf{1}$ onto the orthohgonal space $Lin^{\perp}(\mathbf{1})$ yields zero as demonstrated in Figure . Moreover, it follows that the angle between $\tilde{x}$ and y is still φ.

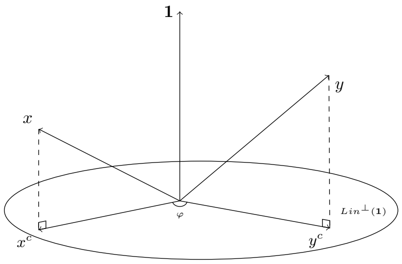


Figure 7: Centred vectors x^c and y^c

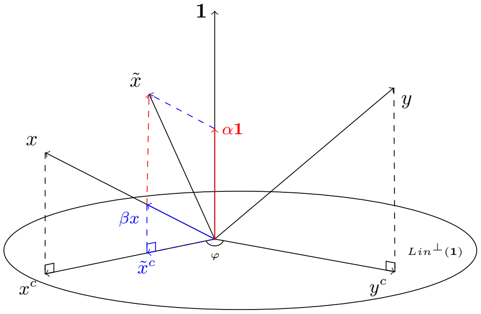


Figure 8: Decomposition and projection of $\tilde{x}$

Finally, putting everything together we finish the proof:

$$\text{sCorr}(x + \alpha\mathbf{1}, y) = \text{sCorr}(x, y)$$

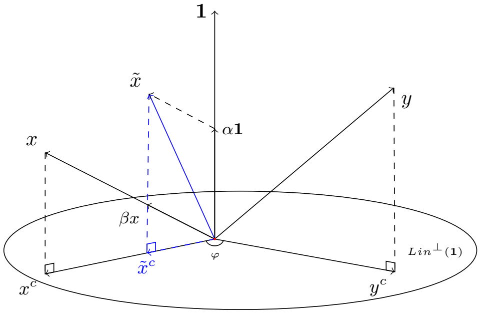


Figure 9: $\text{sCorr}(x + \alpha\mathbf{1}, y) = \text{sCorr}(x, y)$ as the corresponding angles are equal.

□

Sample correlation coefficient in simple linear regression

Theorem 4. *A linear regression model with one explanatory variable and constant term*

$$y = \beta_1 + \beta_2 x + \varepsilon$$

has the property

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

Proof. Firstly, we consider the case when $\hat{\beta}_2 > 0$ and introduce the base picuture of the proof as depicted in Figure 10. It has been shown earlier that the correlation coefficient represents the angle betweem 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 11(a). We denote this space as $Lin^{\perp}(\mathbf{1})$. 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 $\vec{\bar{a}}$.

In order to get the angle between y and $\hat{y}$ we should start with regressing y on $Lin(x, \mathbf{1})$. Then the only thing thing left is to project $\hat{y}$ onto $Lin^{\perp}\mathbf{1}$ 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 when projected onto the orthogonal space. The result of this step is shown in Figure 11(b).

Assuming the underlying relationship between x and y to be

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

where ε_i is an error term the following holds

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

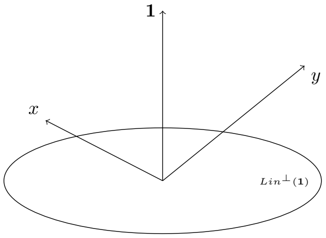
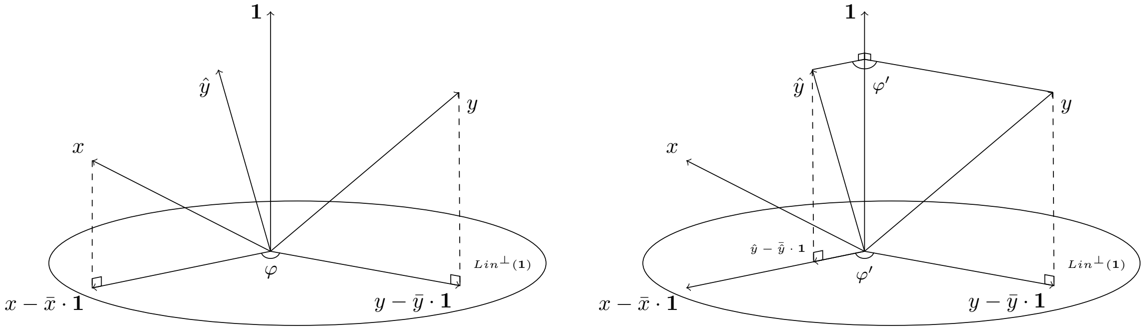


Figure 10: Vectors x, y and $\mathbf{1}$.


$\mathbf{1}$
$\hat{y}$
y
x
φ
$Lin^{\perp}(\mathbf{1})$
$x - \bar{x} \cdot \mathbf{1}$
$y - \bar{y} \cdot \mathbf{1}$
$\mathbf{1}$
φ'
$\hat{y}$
y
x
$\hat{y} - \bar{\hat{y}} \cdot \mathbf{1}$
φ'
$Lin^{\perp}(\mathbf{1})$
$x - \bar{x} \cdot \mathbf{1}$
$y - \bar{y} \cdot \mathbf{1}$

(a)

(b)

Figure 11: (a): ‘Centred’ x and y, i.e., projected onto $Lin^{\perp}(\mathbf{1})$; (b): ‘Centred’ $\hat{y}$, i.e., projected onto $Lin^{\perp}(\mathbf{1})$.

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

Now consider the case when $\hat{\beta}_2 < 0$. Note that the sign of β_1 does not influence the correlation coefficient sign. The only difference is that now $\hat{y}$

is projected onto the span of $x - \bar{x} \cdot \mathbf{1}$ and not on this vector itself while the projections of x and y remain the same. Looking at Figure we deduce that the angle betwween y and $\hat{y}$ is compelement to the angle between x and y. Using trigonometric properties, we simplify $\cos(180° - \varphi) = -\cos\varphi$ which in turn implies sCorr$(x, y) = -$ sCorr$(y, \hat{y})$.

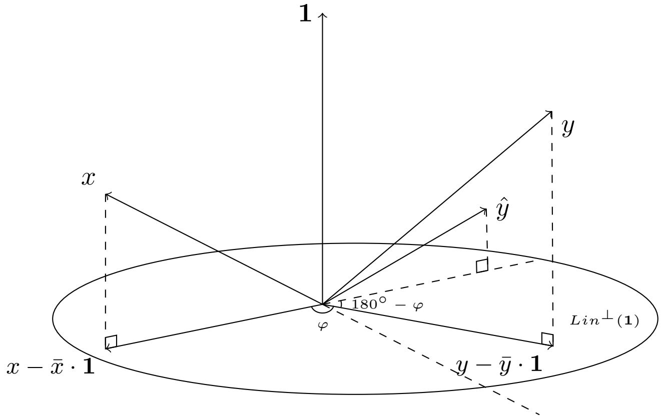

$\mathbf{1}$
y
x
$\hat{y}$
$180° - \varphi$
φ
$Lin^{\perp}(\mathbf{1})$
$x - \bar{x} \cdot \mathbf{1}$
$y - \bar{y} \cdot \mathbf{1}$

Figure 12: Case of $\beta_2 < 0$.

□

RSS + ESS = TSS

Theorem 5. *A linear regression model with n observations and k explanatory variables including a constant unit vector*

$$y = X\beta + \varepsilon$$

has the following property

$$RSS + ESS = TSS$$

where $RSS = \|y - \hat{y}\|_2^2$, $ESS = \|\hat{y} - \bar{y}\|_2^2$, $TSS = \|y - \bar{y}\|_2^2$.

Proof. The proof will be presented for the case of two regressor x and $\mathbf{1}$ in order for the picture to be clear. However, the same logic applies for the case of k regressors.

We start with depicting the vectors $y \in \mathbb{R}^{n-2}$ and $x, \mathbf{1} \in \mathbb{R}^2$. Then we project y onto $Lin(x, \mathbf{1})$ and obtain $\hat{y}$ which is shown in Figure 13(a).

From this picture we can immediately derive $\sqrt{RSS}$ as by definition this is the squared difference between y and $\hat{y}$.

So as to visualize ESS and TSS we first need to visualize vector of averages $\bar{y}$. Geometrically this means projecting a vector onto a line spanned by vector $\mathbf{1}$.

Now we both project y and $\hat{y}$ onto $\mathbf{1}$ and following the definition obtain $\sqrt{TSS}$ as the difference vector $y - \bar{y}$ and $\sqrt{ESS}$ as the vector $\hat{y} - \bar{y}$.

Consider a regresion model with n observations and k explanatory variables including a constant unit vector

$$y = X\beta + \varepsilon$$

The OLS estimator for the vector of coefficients β is

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

and the residual vector is

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

Then we define residual sum of squares (RSS), explained sum of squares (ESS) and total sum of squares (TSS) as follows:

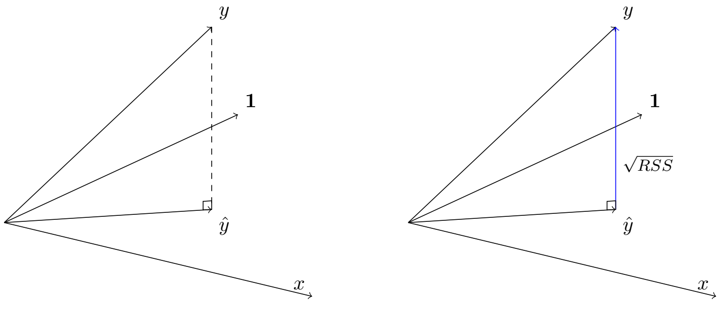
$$\begin{aligned} RSS &= \|y - \hat{y}\|_2^2 \\ ESS &= \|\hat{y} - \bar{y}\|_2^2 \\ TSS &= \|y - \bar{y}\| \end{aligned}$$

Disclosing parentheses and using the fact that $\hat{y}^T y = \hat{y}^T \hat{y}$

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

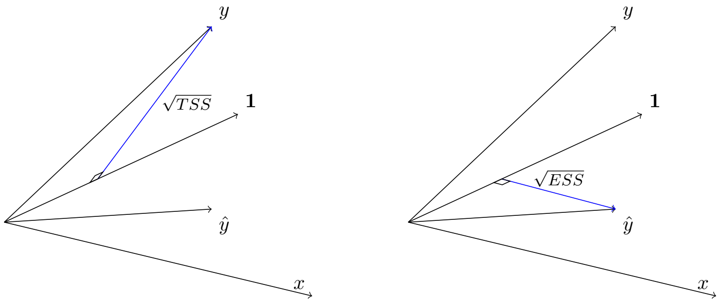
we obtain

$$\begin{aligned} RSS &= y^T y - \hat{y}^T \hat{y} \\ ESS &= \hat{y}^T \hat{y} - \hat{y}^T \bar{y} + \bar{y}^T \bar{y} \\ TSS &= y^T y - 2y^T \bar{y} + \bar{y}^T \bar{y} \end{aligned}$$



(a) (b)

Figure 13: (a): Vectors $y \in \mathbb{R}^{n-2}$ and $\hat{y} \in Lin(x, \mathbf{1})$; (b): Residual sum of squares.



(a) (b)

Figure 14: (a): Total sum of squares; (b): Explained sum of squares.

The final step is to put everything together. Note that since $y - \hat{y}$ is perpendicular to $Lin(x, \mathbf{1})$ it is also perpendicular to $\hat{y} - \bar{y}$ and $\mathbf{1}$ as these vectoros are in $Lin(x, \mathbf{1})$. Then, applying the theorem of three perpendiculars we conclude that the foot of vector $y - \bar{y}$ is the same point as the foot of the vector $\hat{y} - \bar{y}$. Thus, we obtain a right angle triangle and can apply the Pythagorean theorem for the catheti $\sqrt{RSS}$ and $\sqrt{ESS}$ and the hypotenuse $\sqrt{TSS}$:

$$(\sqrt{RSS})^2 + (\sqrt{ESS})^2 = (\sqrt{TSS})^2$$

□

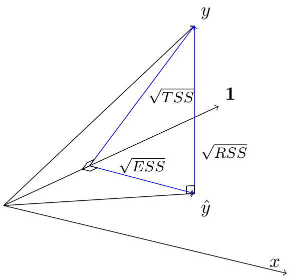


Figure 15: $(\sqrt{RSS})^2 + (\sqrt{ESS})^2 = (\sqrt{TSS})^2$

Determination coefficient

Theorem 6. *A linear regression model with n observations and k explanatory variables including a constant unit vector*

$$y = X\beta + \varepsilon$$

has the following property

$$R^2 = \text{sCorr}^2(y, \hat{y})$$

$$\begin{aligned}
\text{sCorr}^2(y,\hat{y}) &= \left(\frac{\text{sCov}(y,\hat{y})}{\sqrt{\text{sVar}(y)\,\text{sVar}(\hat{y})}}\right)^2 \\
&= \frac{\text{sCov}(y,\hat{y})\,\text{sCov}(y,\hat{y})}{\text{sVar}(y)\,\text{sVar}(\hat{y})} \\
&= \frac{\text{sCov}(\hat{y}+e,\hat{y})\,\text{sCov}(\hat{y}+e,\hat{y})}{\text{sVar}(y)\,\text{sVar}(\hat{y})} \\
&= \frac{(\text{sCov}(\hat{y},\hat{y}) + \text{sCov}(e,\hat{y}))(\text{sCov}(\hat{y},\hat{y}) + \text{sCov}(e,\hat{y}))}{\text{sVar}(y)\,\text{sVar}(\hat{y})} \\
&= \frac{\text{sVar}(\hat{y})\,\text{sVar}(\hat{y})}{\text{sVar}(y)\,\text{sVar}(\hat{y})} = \frac{\text{sVar}(\hat{y})}{\text{sVar}(y)} = \frac{ESS}{TSS} = R^2
\end{aligned}$$

Proof. Proving this theorem geometrically means showing that the determination coefficient can be interpreted as some squared angle which happens to be eqaul to the squared angle betwen y and $\hat{y}$.

Consider Figure 15 from the previous proof. It was shown there that the vectors $y - \bar{y}$, $y - \hat{y}$ and $\hat{y} - \bar{y}$ form a right triangle. Having defined the determination coefficient as

$$R^2 = \frac{ESS}{TSS}$$

we conclude that its geometric interpretaion is

$$R^2 = \frac{ESS}{TSS} = \cos^2 \varphi$$

as shown in Figure .

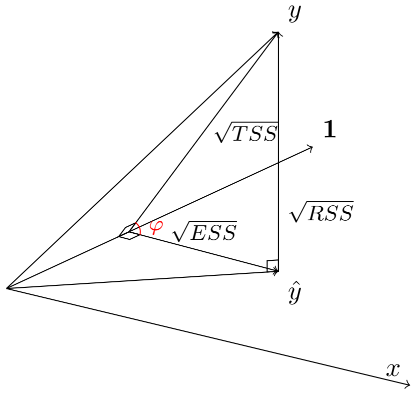


Figure 16: Determination coefficient as squared cos φ

Recall that the sample correlation coefficient two vectors was defined earlier as the angle between these two vectors. Thus, we conclude that $\text{sCorr}(y, \hat{y})$ is the angle between y and $\hat{y}$ which is also eqaul to cos φ. Finally, squaring both sides, we obtain

$$R^2 = \text{sCorr}^2(y, \hat{y})$$

□

Regression line and point of averages

Theorem 7. *In a linear regression model with one explanatory variable and constant term*

$$y = \beta_1 + \beta_2 x + \varepsilon$$

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 $\frac{1}{n}\sum_{i=1}^{n} \hat{y}_i = \frac{1}{n}\sum_{i=1}^{n} y$. In order for the pictures to be more clear the proof will be presented for the case of two regressors.

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

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

Premultiplying both sides by X^T, we obtain:

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

This is a system of equations. The first row of X^T is $\mathbf{1}$ 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

$$X^T\hat{y} = X^Ty$$

The first step is regressing y on $Lin(\mathbf{1}, x)$. As shown in Figure 17(a), we obtain $\hat{y}$ as a linear combination of $\mathbf{1}$ and x. The next step is to regress both y and $\hat{y}$ on $\mathbf{1}$ which results in $\bar{y}$ and $\bar{\hat{y}}$ correspondingly. By the theorem of three perpendiculars, $\bar{y} = \bar{\hat{y}}$ which is shown in Figure 17(b).

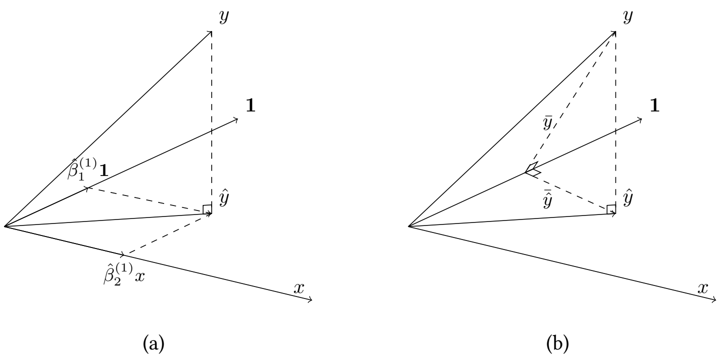


Figure 17: (a): Regression of y on $Lin(\mathbf{1}, x)$; (b): Regression of y and $\hat{y}$ on $\mathbf{1}$.

□

Frisch–Waugh–Lovell theorem

Theorem 8. *Consider regression*

$$y = X_1\beta_1 + X_2\beta_2 + u \tag{1}$$

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, β_1 and β_2 are comfortable, i.e. $k_1 \times 1$ and $k_2 \times 1$ vectors. Consider another regresison

$$M_1 y = M_1 X_2 \beta_2 + M_1 u \tag{2}$$

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 β_2 from regression 1 will be the same as the estimate from regression 2.

From regresison 2 we get the following estimator:

$$\begin{aligned}\hat{\beta}_2 &= ((M_1X_2)^T M_1 X_2)^{-1}(M_1X_2)^T M_1 y\\ &= (X_2^T M_1^T M_1 X_2)^{-1} X_2^T M_1^T M_1 y\\ &= (X_2^T M_1 X_2)^{-1} X_2^T M_1 y\end{aligned}$$

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

$$y = Py + My = X_1\hat{\beta}_1 + X_2\hat{\beta}_2 + My$$

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

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

There are two ways to visualize the proof of the Frisch-Waugh-Lovell theorem using geometric concepts. Both are presented below.

Proof. 1. Consider the following model:

$$y_i = \beta_1 x_i + \beta_2 z_i + u_i \tag{3}$$

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 18(a). Figure 18(b) depicts $Lin(x, z)$.

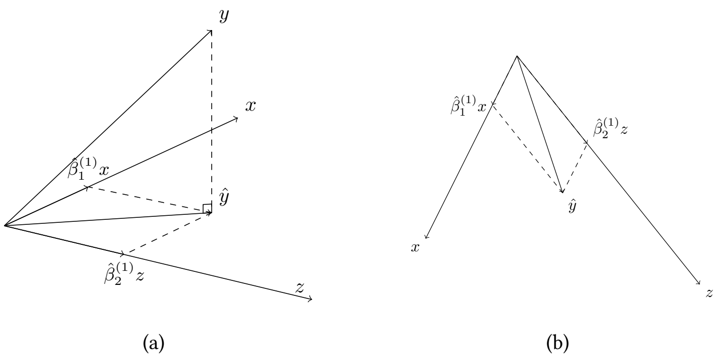


Figure 18: (a): Regression of y on $Lin(x, z)$; (b): $Lin(x, z)$.

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.

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

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

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

Geometric results of these two steps are presented in 19(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 taht obtain $\hat{\beta}_1^{(2)}\tilde{x}$ (see Figure 19(b)).

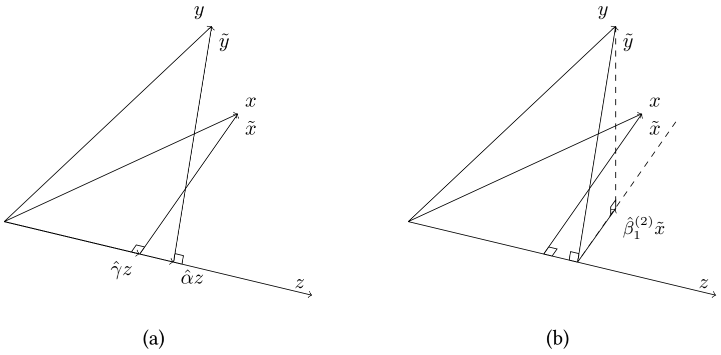
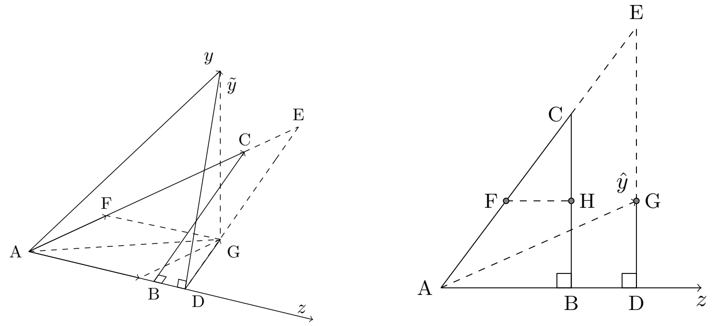


Figure 19: (a): Regression of y on z and of x on z; (b): Translation of $\tilde{x}$.

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



(a)

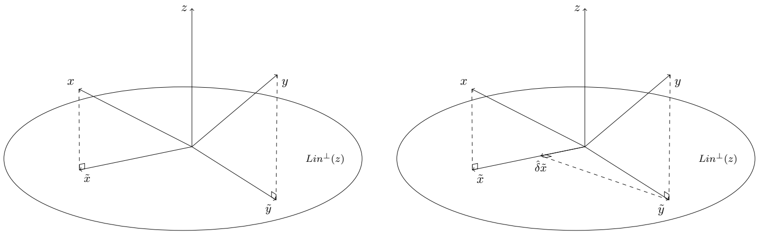
(b)

Figure 20: (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): $Lin(x, z)$.

In Figure 20(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 congruent angles, triangles ABC and FHC are simillar. Then, it follows:

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

2. Alternatively, we could implement a concept close to the partial correlation. In the same model 3 we wiil treat z vector fixed and again consecutively cleanse the x and y variables by projecting them onto the space orthogonal to z, i.e., $Lin^{\perp}(z)$ as demonstrated in Figure 21(a). Then we perform a regression of the 'cleansed' $\tilde{y}$ on the 'cleansed' $\tilde{x}$ (see Figure21(b)).



(a)

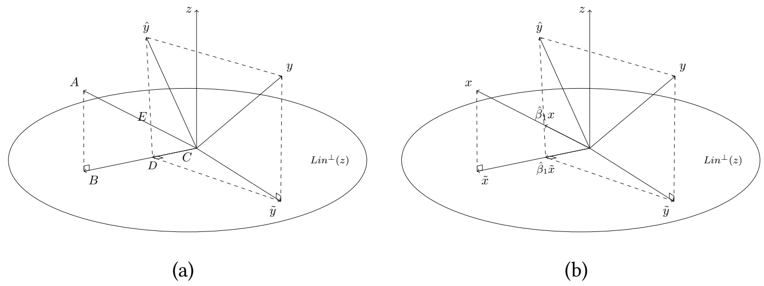
(b)

Figure 21: (a): 'Cleansed' variables $\tilde{x}$ and $\tilde{y}$; (b): 'Cleansed' $\tilde{y}$ regressed on 'cleansed' $\tilde{x}$.

Now we show that the latter regression produces $\hat{\beta}_1$ coefficient which is exatly the coefficient from the 'one-step' regression of original y onto original x and z. Recall that the vector y can be split up into a sum of some multiple of x and some multiple of z. Since the second term is the orthogonal component its projection yields zero. The multiple of x is equal to $\hat{\beta}_1$ by construction.

Assume that the coefficient at $\tilde{x}$ is some unknown variable $\hat{\delta}$. Then consider the similar triangles in the $Lin(x, z)$. From the proportions we obtain:

$$\frac{CE}{CA} = \frac{CD}{CB} \Leftrightarrow \frac{\hat{\beta}_1 x}{x} = \frac{\hat{\delta}\tilde{x}}{\tilde{x}} \Rightarrow \hat{\beta}_1 = \hat{\delta}$$



(a)

(b)

Figure 22: (a): Similar triangles: $\triangle ABC \sim \triangle EDC$; (b): Alternative proof for the Frisch-Waugh-Lovell theorem.

□

Duality of regressors and residuals

The idea of duality is widely used in mathematics. The concept is two apply some transformation twice and get the original object. For example, if $f(a) = 1/a$:

$$x \xrightarrow{f} \frac{1}{x} \xrightarrow{f} \frac{1}{1/x} = x$$

We show that there is duality between regressors and residuals.

Theorem 9. *Let x_i be a $n \times 1$ regressor, u_i — a residual in regression of x_i on all the rest regressors, $i = 1, \ldots, k$. Consider a transformation of a vector v, $f(v) = v/\|v\|^2$. Then applying this transformation on the residuals $u_1, \ldots, u_k$ yields new regressors $v_1, \ldots, v_k$. Performing k regressions of each v_i on all the rest regressors and applying the same transformation to the new residuals results in the original regressors $x_1, \ldots, x_k$.*

Proof. We start with 2-dimensional case with two regressors, and discuss the case of spaces of higher dimensions later.

As stated in the theorem we need to keep the measure of the lengths of the regressors. In order to do this we choose a basis in $\mathbb{R}^2$ in such a way that

$$x_1 = \lambda_1 e_1, \quad \|e_1\| = 1$$
$$x_2 = \lambda_2 e_2, \quad \|e_2\| = 1$$

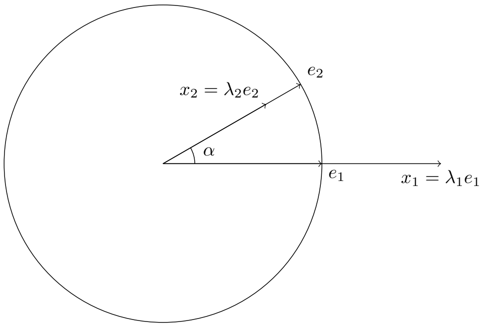


Figure 23: Two regressors in the unit circle.

where $\lambda_1, \lambda_2 \in \mathbb{R}$.

Then we perform two regressions

$$x_1 = \beta_1 x_2 + u_1$$
$$x_2 = \beta_2 x_1 + u_2$$

and get the residuals $\hat{u}_1$, $\hat{u}_2$. Being orthogonal to x_2 and x_1 correspondingly, they can be written as follows

$$\hat{u}_1 = \sin\alpha \cdot \lambda_1 \tilde{e}_1, \quad \|\tilde{e}_1\| = 1$$
$$\hat{u}_2 = \sin\alpha \cdot \lambda_2 \tilde{e}_2, \quad \|\tilde{e}_2\| = 1$$

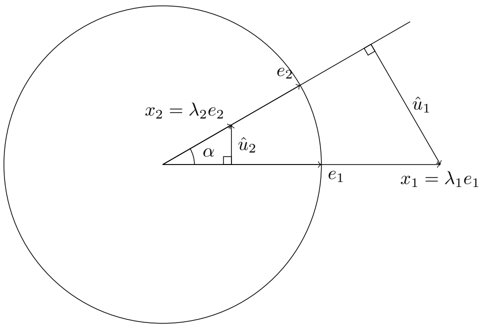


Figure 24: Residuals $\hat{u}_1$ and $\hat{u}_2$

where $\tilde{e}_1 \perp e_2$ and $\tilde{e}_2 \perp e_1$.

For convinence we translate all the vectors x_1, x_2, $\hat{u}_1$, $\hat{u}_2$ to the origin of the unit circle as shown in Figure 26(a) and after that we invert them.

In order to illustrate inversion consider an example with an arbitrary vector a. Knowing its length, the aim is to find such an orthogonal vector $\tilde{a}$ that the product $\|a\|^2 \cdot \|\tilde{a}\|^2 = 1$. In other words, we need to find an edge of rectangle with area equal to 1. Solving for $\tilde{a}$, we obtain the length of the inverted vector a. The only thing left is to rotate this inverted vector back to get a vector $\tilde{\tilde{a}}$ which satisfies both

$$\|\tilde{\tilde{a}}\|^2 = \frac{1}{\|a\|^2}$$
$$\cos(a, \tilde{\tilde{a}}) = 1$$

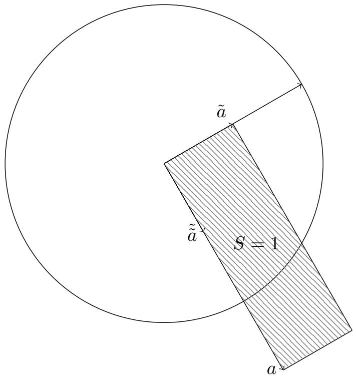


Figure 25: Example of inversion for vector a.

Having applied the inversion to $\hat{u}_1$, $\hat{u}_2$, we obtained new vectors y_1, y_2. Morevover, there is an algebraic expression for them in terms of rotated basis $\tilde{e}_1$, $\tilde{e}_2$:

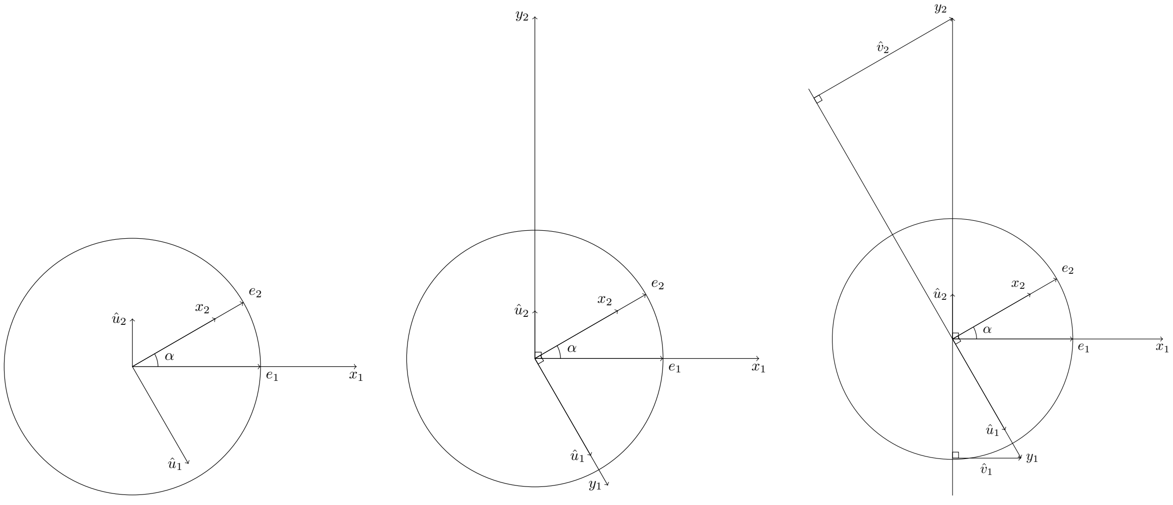
$$\hat{u}_1 = \sin\alpha \cdot \lambda_1 \tilde{e}_1 \Rightarrow y_1 = \frac{1}{\sin\alpha \cdot \lambda_1}\tilde{e}_1$$
$$\hat{u}_2 = \sin\alpha \cdot \lambda_2 \tilde{e}_2 \Rightarrow y_2 = \frac{1}{\sin\alpha \cdot \lambda_2}\tilde{e}_2$$

The transformation stated in the theorem is $f(v) = v/\|v\|^2$. Generally speaking, $g(v) = v/(c \cdot \|v\|^2)$ where $c \in \mathbb{R}$ would also work.

$$v \xrightarrow{g} \frac{v}{c \cdot \|v\|^2} = w \xrightarrow{g}$$
$$\frac{w}{c \cdot \|w\|^2} = \frac{\frac{v}{c \cdot \|v\|^2}}{c \frac{\|v\|^2}{c^2 \|v\|^4}} = v$$

Next, we perform another two regressions:

$$y_1 = \gamma_1 y_2 + v_1$$
$$y_1 = \gamma_2 y_1 + v_2$$



(a) (b) (c)

Figure 26: (a): Residuals translated to the orgin of the unit circle; (b): Regressors v_1, v_2 obtained from inversion of the residuals $\hat{u}_1$, $\hat{u}_2$; (c): Regressions of v_1 onto v_2 and of v_2 onto v_1.

There two things to notice about the new residuals $\hat{v}_1$, $\hat{v}_2$. First, $\hat{v}_1$ is perpendicular to the line spanned by $\tilde{e}_2$. Similarly, $\hat{v}_2$ is perpendicular to the line spanned by $\tilde{e}_1$. This means, that they are parallel to e_1, e_2 correspondingly and once translated, they can be expressed as a multiple of x_1, x_2.

Second, we can find the lengths of these new residuals from the right triangles depicted in Figure 26(c):

$$\|\hat{v}_1\| = \sin\alpha \cdot \|y_2\| = \sin\alpha \cdot \left\| \frac{1}{\sin\alpha \cdot \lambda_1} \tilde{e}_1 \right\| = \frac{1}{\lambda_1}$$
$$\|\hat{v}_2\| = \sin\alpha \cdot \|y_1\| = \sin\alpha \cdot \left\| \frac{1}{\sin\alpha \cdot \lambda_2} \tilde{e}_2 \right\| = \frac{1}{\lambda_2}$$

Thus, when translated to the origin, the new resiuduals can be rewritten as

$$\hat{v}_1 = \frac{1}{\lambda_1} e_1$$
$$\hat{v}_2 = \frac{1}{\lambda_2} e_2$$

The last step is to invert $\hat{v}_1$, $\hat{v}_2$. Following the same procedure as described above, we finally get the desired result:

$$\hat{v}_1 = \frac{1}{\lambda_1} e_1 \to \lambda_1 e_1 = x_1$$
$$\hat{v}_2 = \frac{1}{\lambda_2} e_2 \to \lambda_2 e_2 = x_2$$

□

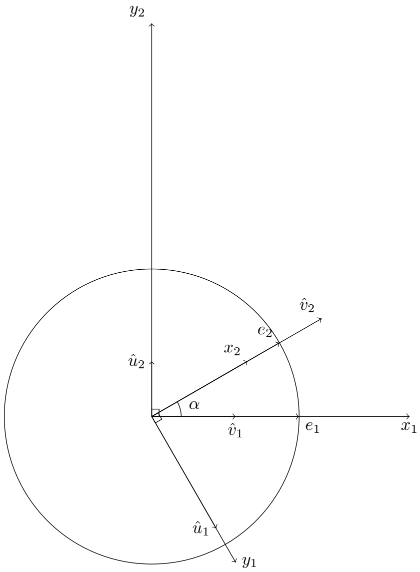


Figure 27: New residuals translated to the origin of the unit circle.

Partial correlation

Definition of partial correlation

Partial correlation can be defined in two ways. We will provide both definitions and show their equivalence.

Partial correlation is a measure of degree of dependence between two random variables while controlling for the effect of other random variables:

$$\text{pCorr}(X, Y; Z) = \frac{\text{pCov}(X, Y; Z)}{\sqrt{\text{pVar}(X; Z)\,\text{pVar}(Y; Z)}}$$

where $\text{pVar}(X; Z) = \text{Var}(X - \alpha Z)$, α is such a constant that $\text{Cov}(X - \alpha Z, Z) = 0$, and $\text{pCov}(X, Y; Z) = \text{Cov}(X - \alpha Z, Y - \beta Z)$, α, β are such constants that $\text{Cov}(X - \alpha Z, Z) = \text{Cov}(Y - \beta Z, Z) = 0$.

Definition 1. Partial correlation between random variables X and Y holding random variable Z fixed is the correlation coefficient between the residuals in regression of X onto Z and the residuals in regression of Y onto Z.

Firstly, we project random variable X onto Z, which yields $\text{E}(X|Z)$. The residuals in this regression are $X - \text{E}(X|Z)$ — a vector in $Lin^{\perp}(Z)$. We will call this variable 'cleansed' and label it as $\widetilde{X}$. Applying the same procedure for Y yields 'cleansed' variable $\widetilde{Y} = Y - \text{E}(Y|Z) \in Lin^{\perp}(Z)$. The angle between $\widetilde{X}$ and $\widetilde{Y}$ (φ in Figure) is the correaletion coefficient between these 'cleansed' random variables and the partial correlaiton between the original ones.

Let us define $\widetilde{X}$ and $\widetilde{Y}$ as

$$\widetilde{X} = \alpha_{XY}\widetilde{Y} + \tilde{u}_{XY}, \widetilde{X} \perp Z$$
$$\widetilde{Y} = \alpha_{YX}\widetilde{X} + \tilde{u}_{YX}, \widetilde{Y} \perp Z$$

Then assuming that the error term u_{XY} is uncorrelated with $\widetilde{Y}$, we obtain:

$$\text{Cov}(\widetilde{Y}, \widetilde{X} - \alpha_{XY}\widetilde{Y}) = 0$$
$$\text{Cov}(\widetilde{Y}, \widetilde{X}) - \alpha_{XY}\,\text{Cov}(\widetilde{Y}, \widetilde{Y}) = 0$$
$$\alpha_{XY} = \frac{\text{Cov}(\widetilde{Y}, \widetilde{X})}{\text{Var}(\widetilde{Y})}$$

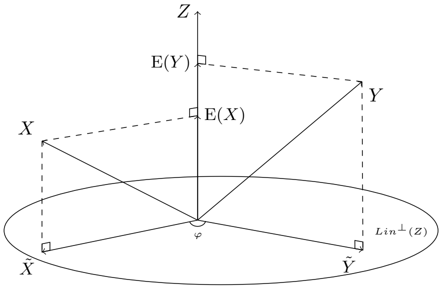


Figure 28: Partial correlation between X and Y while Z is fixed.

Definition 2. Partial correlaiton between random variables X and Y holding random variable Z fixed is the geometric mean between the coefficient β_{XY} in regression

$$X = \beta_{XY} Y + \beta_{XZ} Z + u_X$$

and the coefficinent β_{YX} in regression

$$Y = \beta_{YX} X + \beta_{YZ} Z + u_Y$$

Partial correlation has the same sign as the coefficients β_{XY} and β_{YX}.

Following the definition, we need to start with regressing variable X onto Y and Z. Then, the vector we obtain $\hat{X} = \mathrm{E}(X|Y, Z)$ can be broken up into the sum of $\beta_{XY} Y$ and $\beta_{XZ} Z$. Projecting $\beta_{XY} Y$ onto $Lin^{\perp}(Z)$ results in a vector $\alpha_{XY} \widetilde{Y}$ where $\widetilde{Y} = Y - \mathrm{E}(Y|Z)$ is the projection of Y onto $Lin^{\perp}(Z)$.

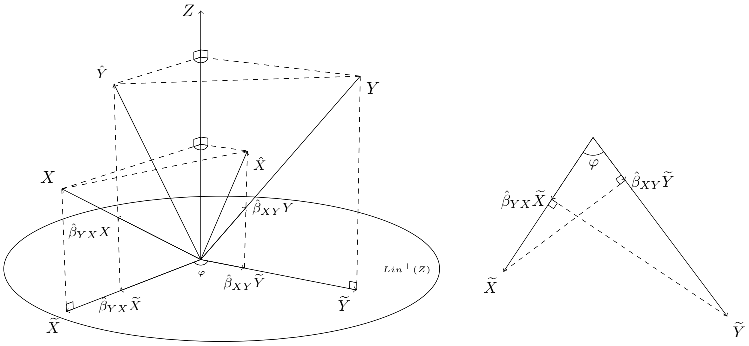
By the properties of similar triangles

$$\frac{\beta_{XY} Y}{Y} = \frac{\alpha_{XY} \widetilde{Y}}{\widetilde{Y}} \Leftrightarrow \beta_{XY} = \alpha_{XY}$$

Multiplying these coefficients, we get the final result:

$$\alpha_{XY} \cdot \alpha_{XY} = \frac{\mathrm{Cov}^2(\widetilde{Y}, \widetilde{X})}{\mathrm{Var}(\widetilde{Y})\,\mathrm{Var})\widetilde{X}}$$
$$= \mathrm{Corr}^2(\widetilde{X}, \widetilde{Y}) = \mathrm{pCorr}^2(X, Y; Z)$$

In the same way we perform a regression of Y onto X and Z and repeat the same steps for X. Finally, we get the whole picture:



(a) (b)

Figure 29: (a): Alternative definition of the partial correlation; (b): $Lin^{\perp}(Z)$.

Having plotted $Lin^{\perp}(Z)$, now we can express $\cos\varphi$ in terms of β_{XY} and β_{YX}:

$$\cos\varphi = \frac{|\beta_{XY}\widetilde{Y}|}{|\widetilde{X}|} = |\beta_{XY}|$$

$$\cos\varphi = \frac{|\beta_{YX}\widetilde{X}|}{|\widetilde{Y}|} = |\beta_{YX}|$$

$$\cos^2\varphi = |\beta_{XY}\beta_{YX}| \overset{\beta_{XY}\beta_{YX}>0}{=} \beta_{XY}\beta_{YX}$$

Recall that the angle φ can be interpreted as the correlation between $\widetilde{X}$ and $\widetilde{Y}$. These random variables are constructed in such a way that both of them are uncorrelated with Z. Thus, it follows that

$$\cos^2\varphi = \operatorname{Corr}^2(\widetilde{X}, \widetilde{Y}) = \operatorname{pCorr}^2(X, Y; Z) = \beta_{XY}\beta_{YX}$$

Partial correlation as correlaiton between residuals

Theorem 10. *Partial correlaiton between X and Y holding Z fixed is the negative correlaiton coefficient betwenn the residuals u in the regeression model*

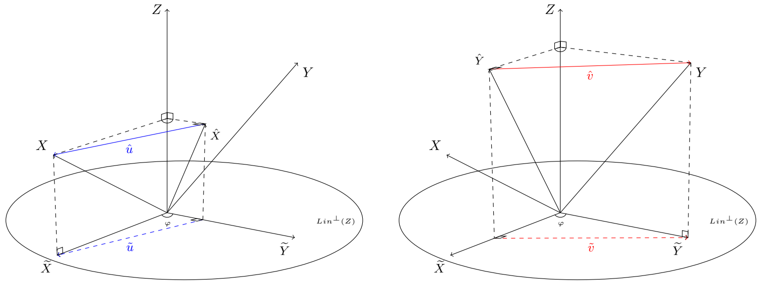
$$X = \alpha_1 Y + \alpha_2 Z + u$$

and the residuals v in the model

$$Y = \beta_1 X + \beta_2 Z + v$$

Proof. The first step is to find the residuals in the mentioned regresisions. For example, in order to get u we regress X onto $Lin(Y, Z)$ which results in $\hat{X} = X - \operatorname{E}(X|Y, Z)$. Then we take the difference $X - \hat{X} = u$ and project it as well as X itself onto $Lin^{\perp}(Z)$ as demonstrated in Figure 30(a). We denote the result as $\tilde{u}$ and $\widetilde{X}$ respectively.

Figure 30(b) shows the same steps for obtainig $\tilde{v}$ and $\widetilde{Y}$.



(a)

(b)

Figure 30: (a): $\hat{u}$ form regression of x onto Y and Z, $\hat{u}$ projected; (b): $\hat{v}$ from regression of Y onto X and Z, $\hat{v}$ projected.

After putting these figures together, we need to measure the angle between the $\tilde{u}$ and $\tilde{v}$. Translating the $\tilde{v}$ vector to the orgin of $\tilde{x}$ as shown in Figure 31(b), we conclude that the desired angle is the bigger one of the vertical angles. Hence, we can derive it from the property of quadrilateral by

Let us define cleansed X and Y first as

$$\widetilde{X} = \alpha_1\widetilde{Y} + \tilde{u}, \widetilde{X} \perp Z$$
$$\widetilde{Y} = \beta_1\widetilde{X} + \tilde{v}, \widetilde{Y} \perp Z$$

Then

$$\alpha_1 = \frac{\operatorname{Cov}(\widetilde{X}, \widetilde{Y})}{\operatorname{Var}(\widetilde{Y})}$$
$$\beta_1 = \frac{\operatorname{Cov}(\widetilde{X}, \widetilde{Y})}{\operatorname{Var}(\widetilde{X})}$$

Substituting these into $\operatorname{Cov}(\tilde{u}, \tilde{v})$, we obtain:

$$\operatorname{Cov}(\tilde{u}, \tilde{v}) =$$
$$\operatorname{Cov}\left(\widetilde{X} - \frac{\operatorname{Cov}(\widetilde{X}, \widetilde{Y})}{\operatorname{Var}(\widetilde{Y})}\widetilde{Y}, \widetilde{Y} - \frac{\operatorname{Cov}(\widetilde{X}, \widetilde{Y})}{\operatorname{Var}(\widetilde{X})}\widetilde{X}\right)$$
$$= \operatorname{Cov}(\widetilde{X}, \widetilde{Y}) - \operatorname{Cov}(\widetilde{X}, \widetilde{Y})$$
$$- \operatorname{Cov}(\widetilde{X}, \widetilde{Y}) + \frac{\operatorname{Cov}^3(\widetilde{X}, \widetilde{Y})}{\operatorname{Var}(\widetilde{X})\operatorname{Var}(\widetilde{Y})}$$
$$= -\operatorname{Cov}(\widetilde{X}, \widetilde{Y})\left(1 - \frac{\operatorname{Cov}^2(\widetilde{X}, \widetilde{Y})}{\operatorname{Var}(\widetilde{X})\operatorname{Var}(\widetilde{Y})}\right)$$

Next, we deal with variances of $\tilde{u}$ and $\tilde{v}$:

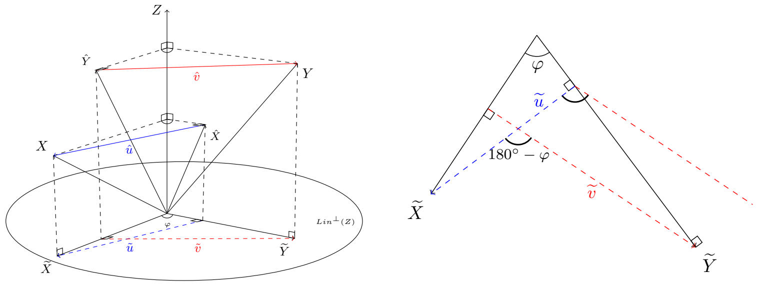
$$\operatorname{Var}(\tilde{u}) = \operatorname{Var}(\widetilde{X}) - \frac{\operatorname{Cov}^2(\widetilde{X}, \widetilde{Y})}{\operatorname{Var}^2(\widetilde{Y})}\operatorname{Var}(\widetilde{Y})$$
$$- 2\operatorname{Cov}(\widetilde{X}, \widetilde{Y})\frac{\operatorname{Cov}(\widetilde{X}, \widetilde{Y})}{\operatorname{Var}(\widetilde{Y})}$$
$$= \operatorname{Var}(\widetilde{X})\left(1 - \frac{\operatorname{Cov}^2(\widetilde{X}, \widetilde{Y}}{\operatorname{Var}(\widetilde{X})\operatorname{Var}(\widetilde{Y})}\right)$$
$$\operatorname{Var}(\tilde{v}) = \operatorname{Var}(\widetilde{Y})\left(1 - \frac{\operatorname{Cov}^2(\widetilde{X}, \widetilde{Y}}{\operatorname{Var}(\widetilde{X})\operatorname{Var}(\widetilde{Y})}\right)$$

Now we can write out $\operatorname{Corr}(\tilde{u}, \tilde{v})$:

$$\operatorname{Corr}(\tilde{u}, \tilde{v}) =$$
$$= -\frac{\operatorname{Cov}(\widetilde{X}, \widetilde{Y})\left(1 - \frac{\operatorname{Cov}^2(\widetilde{X}, \widetilde{Y})}{\operatorname{Var}(\widetilde{X})\operatorname{Var}(\widetilde{Y})}\right)}{\sqrt{\operatorname{Var}(\widetilde{X})\operatorname{Var}(\widetilde{Y})\left(1 - \frac{\operatorname{Cov}^2(\widetilde{X}, \widetilde{Y})}{\operatorname{Var}(\widetilde{X})\operatorname{Var}(\widetilde{Y})}\right)^2}}$$
$$= -\operatorname{Corr}(\widetilde{X}, \widetilde{Y})$$

substracrtion all the known angles from 360°. Thus, the desired one is equal to $180° - \varphi$.

$$\begin{aligned}\cos(\varphi) &= -\cos(180° - \varphi)\\ \mathrm{Corr}(\widetilde{X}, \widetilde{Y}) &= -\,\mathrm{Corr}(\tilde{u}, \tilde{v})\\ \mathrm{pCorr}(X, Y; Z) &= -\,\mathrm{Corr}(u, v)\end{aligned}$$



(a) (b)

Figure 31: (a): The residuals of both regressions; (b): $Lin^{\perp}(Z)$.

□

Probability distributions

Normal

By contrast with substantial majority of books, the univariate normal distribution can be derived form the multivariate normal distribution. In this section we show how to obtain the univariate normal from the bivariate.

Assume there are gas molecules moving chaotically on a plane and we can measure the velocity vector of one of them. We denote this vector as $V = \begin{pmatrix} X \\ Y \end{pmatrix}$ where X and Y satnd for the horizontal and vertical components respectively. Assume additionally that

1. the joint distribution finction $f(x, y)$ does not depend on the vector $(X, Y)^T$ direction but depends on its length only,
2. the orthogonal components of the velocity vector are independent.

Theorem 11. *The only way the assumptions (1) and (2) are satisfied is when $X \sim \mathcal{N}(0, \sigma^2)$, $Y \sim \mathcal{N}(0, \sigma^2)$ and X, Y are independent.*

Proof. First of all, consider a vector $V' = \begin{pmatrix} -Y \\ X \end{pmatrix}$, i.e., the original vector V rotated 90° counterclock-wise. By the assumption (1), this operation did not change the distribution of V. Hence, $V' \sim V$ which implies $-Y \sim X$ and

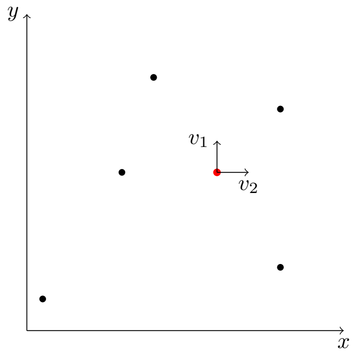


Figure 32: Black dots represent the gas molecules. The red dot stands for the one we catch. Its speed along the horizontal axis is v_1, i.e., the first component of the velocity vector, and its speed along the vertical axis is v_2.

It is more common to introduce a standard normal distribution in terms of its PDF

Definition 3. A continuous random variable ξ has a standard normal distribution if its PMF is given by

$$f_\xi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right).$$

After that multivariate normal is defined.

Definition 4. Let $\xi_i \overset{iid}{\sim} \mathcal{N}(0, 1)$ then $\xi \sim \mathcal{N}(\vec{0}, I)$ where $\xi = \begin{pmatrix} \xi_1 \\ \vdots \\ \xi_n \end{pmatrix}$, I is $n \times n$ identity matrix, and its PMF is

$$f_\xi(x_1, \ldots, x_n) = \frac{1}{(\sqrt{2\pi})^n} \exp\left(-\frac{x_1^2 + \ldots + x_n^2}{2}\right).$$

And finally, location-scale transformations are applied.

$X \sim Y$. It follows that

$$\begin{cases} \mathrm{E}(-Y) = \mathrm{E}(X) \\ \mathrm{E}(X) = \mathrm{E}(Y) \end{cases}$$

which holds for $\mathrm{E}(X) = \mathrm{E}(Y) = 0$ only.

Likewise, $\mathrm{Var}(X) = \mathrm{Var}(Y)$ and here we additionally assume without loss of generality that $\mathrm{Var}(X) = \mathrm{Var}(Y) = 1$.

Next, we introduce the angle between V and the horizontal axis U and the length of the velocity vector $R = \sqrt{X^2 + Y^2}$. Obviously, $X = R \cos U$ and $Y = R \sin U$. Note that since the joint distribution of X and Y depends only on the length of vector V the distribution function of U can only be constant, thus $U \sim Unif(0, 2\pi)$.

Applying assumption (1) again, we conclude that the joint distribution function can be written as a function of the length of the velocity vector, or equivalently of the length squared:

$$f(x, y) = h(x^2 + y^2).$$

By the assumption (2), orthogonal components of V are independent. Hence, the joint PMF can be decomposed to the product of marginal ones:

$$f(x, y) = f(x)f(y) = g(x^2)g(y^2)$$

where the latter equality was written for convenience. Putting everything together, we obtain

$$h(x^2 + y^2) = g(x^2)g(y^2).$$

Next, we take the derivative of both sides with respect to y^2 and then substitute $y^2 = 0$ to get a constant k:

$$\begin{aligned} h'(x^2 + y^2) &= g(x^2)g'(y^2) \\ h'(x^2) &= g(x^2)g'(0) \\ h'(x^2) &= k \cdot g(x^2) \end{aligned}$$

Solving the differential equation, we obtain

$$h(x^2) = ce^{kx^2}, \quad c \in \mathbb{R}.$$

So the joint PMF can be written as follows:

$$f(x, y) = h(x^2 + y^2) = ce^{k(x^2+y^2}$$

and due to independece of X and Y the PMF of X is

$$f(x) = \sqrt{c}e^{kx^2}.$$

In oreder to find the constant k, we need to solve $\mathrm{E}(X^2) = 1$. Computing the integral we obtain $k = -\frac{1}{2}$.

In order to obtain k we computed

$$\begin{aligned} \mathrm{E}(X^2) &= \int_{-\infty}^{\infty} x^2 \sqrt{c} e^{kx^2} dx \\ &= x \cdot \sqrt{c} e^{kx^2} \cdot \frac{k}{2}\Big|_{-\infty}^{\infty} - \int_{-\infty}^{\infty} \sqrt{c} e^{kx^2} \cdot \frac{1}{2k} dx \\ &= -\frac{1}{2k} \int_{-\infty}^{\infty} \sqrt{c} e^{kx^2} \\ &= -\frac{1}{2k} \cdot 1 \\ &= 1. \end{aligned}$$

Finally, we need to normalize $f(x) = ce^{-\frac{x^2+y^2}{2}}$ so as to obtain c. Again, computing another integral, we conclude that $c = (2\pi)^{-1}$ which finishes the proof. □

In order to obtain c we computed:

$$\begin{aligned} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-\frac{x^2+y^2}{2}} dx dy &= \int_0^{2\pi} \int_0^{\infty} e^{-\frac{r^2}{2}} r dr d\theta \\ &= \int_0^{2\pi} \left(\int_0^{\infty} e^{-u} du \right) d\theta \\ &= \int_0^{2\pi} 1 d\theta \\ &= 2\pi. \end{aligned}$$

Notice, that any other $\mathcal{N}(\mu, \sigma^2)$ can be obtained by applying location-scale transformations.

The theorem can be generalized to the n-dimensional case.

Theorem 12. *The vector* $z = \begin{pmatrix} z_1 \\ \vdots \\ z_n \end{pmatrix}$ *folllows the standard multivariate normal distribution and its components are independent if and only if*

1. *f(z) depends on* $|z|$ *only,*
2. *the projections of vector* z *onto the orthogonal subspaces* A *and* B *in* $\mathbb{R}^n$ *are independent.*

Chi-squared

Theorem 13. *Consider a random vector* $z \in \mathbb{R}^n$ *which components are independent and follow standard normal distribution,* $z_i \sim \mathcal{N}(0, 1)$*. Consider also a fixed* k*-dimensional subspace* L *in* $\mathbb{R}^n$*. Let the projection of vector* z *onto the subspace* L *be* $\hat{z}$ *and its length squared* Q

$$Q = \|\hat{z}\|^2 = \langle \hat{z}, \hat{z} \rangle = \hat{z}^T \hat{z}$$

Then Q *follows the chi-squared distribution with* k *degrees of freedom.*

Let $z_i \overset{iid}{\sim} \mathcal{N}(0, 1)$. Then Q follows the chi-squared distribution with k degrees of freedom if it can be written as

$$Q = z_1^2 + z_2^2 + \ldots + z_k^2.$$

This definition is a particular case of the geometric one. Consider projecting a vector $z = (z_1, z_2, \ldots, z_n)$ form $\mathbb{R}^n$ onto the k-dimensional subspace S of vectors which first k coordinates are arbitrary and all the rest are zeros. As a result we would get

$$\hat{z} = (z_1, z_2, \ldots, z_k, 0, \ldots, 0).$$

Squaring the length of the projection, we obtain

$$Q = \|\hat{z}\| = z_1^2 + z_2^2 + \ldots + z_k^2.$$

Proof. First, it can be shown that the projected vector $\hat{z}$ is the original vector z multiplied by the projection matrix $H = X(X^T X)^{-1} X^T$ where the columns of X are fixed linearly independent vectors $x_1, \ldots, x_k$ in L or equivalently $colX = Lin(x_1, \ldots, x_k)$. This matrix is also often referred to as 'hat-matrix'. Then the statement in the theorem can be rewritten as follows:

$$\hat{z}^T \hat{z} = (Hz)^T Hz = z^T H^T Hz = z^T H^2 z = z^T Hz,$$

applying the idempotence property in the last step.

Another nice property of the hat-mattix is symmetry. Thus, it can be decomposed as

$$H = PDP^T,$$

where we choose the vectors of matrix P to be unit and orthogonal, and $D = diag(\lambda_1, \ldots, \lambda_n)$ where λ_i is an eigenvalue of H.

Since $H^2 = H$ the eigenvalues are either 0 or 1. Recall that H projects a vector onto $colX$. Then for any x_i, $i = 1, \ldots, k$, $Hx_i = x_i \cdot 1$ since any x_i is already in $colX$. This implies that $\lambda_1 = \ldots = \lambda_k = 1$. There are also $n - k$ vectors in the subspace orthogonal to $colX$. So for any x_i, $i = k + 1, \ldots, n$, the orthogonal projection yields zero. We conclude that $\lambda_{k+1} = \ldots = \lambda_n = 0$.

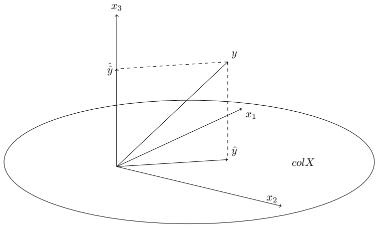


Figure 33: Consider a 3-dimensional example, $colX = Lin(x_1, x_2)$ and $col^{\perp}X = Lin(x_3)$. $Hx_1 = x_1$ and $Hx_2 = x_2$ since they are in $colX$. However, $Hx_3 = 0$ as $x_3 \perp colX$. Projecting an arbitrary vector onto $colX$ yileds $Hy = \hat{y} \in Lin(x_1, x_2)$ while projecting onto $col^{\perp}X$ results in $(I - H)y = \hat{\hat{y}} \in Lin(x_3)$.

Rewritting the theroem statement further, we obtain

$$z^T H z = z^T P D P^T z = (P^T z)^T D (P^T z) = \tilde{z}^T D \tilde{z} = \tilde{z}_1^2 + \ldots + \tilde{z}_k^2.$$

Now we explore $\tilde{z}$ given $z \sim \mathcal{N}(0, I)$:

$$\begin{aligned} &\tilde{z} = P^T z \\ &\mathrm{E}(\tilde{z}) = \mathrm{E}(P^T z) = P^T \mathrm{E}(z) = 0 \\ &\mathrm{Var}(\tilde{z}) = \mathrm{Var}(P^T z) = P^T \mathrm{Var}(z) (P^T)^T = P^T P = I \end{aligned}$$

So we conclude that $\tilde{z}_1^2 + \ldots + \tilde{z}_k^2 \sim \chi_k^2$.

□

Student's

A continuous random variable T has Student's distribution with k degrees of freedom if it can be expressed as

$$T = \frac{Z}{\gamma_k / k},$$

where $Z \sim \mathcal{N}(0, 1)$, $\gamma_k \sim \chi_k^2$ and Z, γ_k are independent.

Definition 5. Let $z = \begin{pmatrix} z_0 \\ z_1 \\ \vdots \\ z_k \end{pmatrix}$ where $z_i \sim \mathcal{N}(0, \sigma^2), i = 0, \ldots, k$ and are independent. Let L_1 be 1-dimensional subspace in $\mathbb{R}^{k+1}$, L_2 an orthogonal to L_1 subspace. Then

$$T = \frac{\|H_1 z\|}{\|H_2 z\| / \sqrt{dim L_2}}$$

where $\|H_1 z\|$ is the length of projection onto L_1 of the vector z, $\|H_2 z\|$ — the length of the projection onto L_2, follows Student's distribution with $dim L_2$ degrees of freedom.

Let us choose two orthogonal subspaces: one-dimensional L_1 and k-dimensional L_2.

Previously we showed that, the squared length of projection follows the chi-squared distribution with the degrees of freedom equal to the dimension onto which the vector was projected. Thus, $\|H_1 z\|^2 \sim \chi_1^2$ and $\|H_2 z\|^2 \sim \chi_k^2$. Now we can express T^2 as a ratio of the per-dimension lengths squared:

$$T^2 = \frac{\|H_1 z\|^2}{\|H_2 z\|^2 / dim L_2}$$

Taking the square root of both sides, we obtain:

$$T = \frac{\|H_1 z\|}{\|H_2 z\| / \sqrt{dim L_2}} = \mathrm{tg}\, \varphi$$

The latter equlity can be illustarted with a 3-dimensional example (see Figure 35).

t-test

In a simple linear regression model

$$y = \beta_1 + \beta_2 x + \varepsilon$$

the adjusted t-value $\frac{t}{\sqrt{n-2}}$ when $H_0 : \beta_2 = 0$ is tested can be expreesed in terms of the angle between y and $\hat{y}$ φ and is equal to ctg φ.

Recall that the t-statistic is defined in the following way:

$$t = \frac{\hat{\beta} - \beta}{s.e.(\hat{\beta})}$$

Adjusting this formula for the null hypothesis $H_0 : \beta_2 = 0$, we obtain

$$t = \frac{\hat{\beta}_2}{s.e.(\hat{\beta}_2)} \tag{6}$$

Then, we need to express $s.e.(\hat{\beta}_2)$ in terms of vectors which can be plotted. From standard OLS procedure it follows that

$$\text{Var}(\hat{\beta}_2) = \frac{\sigma}{\sum\limits_{i=1}^{n}(x_i - \bar{x})^2} \tag{7}$$

Since actual σ is unknown the estimator will be used instead:

$$\hat{\sigma}^2 = \frac{RSS}{n-2} \tag{8}$$

Substituting (7) and (8) into (6) divided by $\sqrt{n-2}$, we obtain

$$\frac{t}{\sqrt{n-2}} = \frac{\hat{\beta}_2}{\sqrt{n-2}\widehat{s.e.}(\hat{\beta}_2)} = \frac{\hat{\beta}_2}{\sqrt{n-2}\frac{\hat{\sigma}}{\sqrt{\sum\limits_{i=1}^{n}(x_i-\bar{x})^2}}}$$

$$= \frac{\hat{\beta}_2\sqrt{\sum\limits_{i=1}^{n}(x_i - \bar{x})^2}}{\sqrt{n-2}\frac{\sqrt{\sum\limits_{i=1}^{n}(y_i-\hat{y}_i)^2}}{\sqrt{n-2}}} = \frac{\hat{\beta}_2|x^c|}{\sqrt{RSS}}$$

where $|x^c| = \sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}$ is the length of the centred vector x.

Now the result can be demonstrated visually. Again we will project x and y vectors onto the $Lin^{\perp}(\mathbf{1})$ so as to get their centred versions x^c and y^c. Then, we perform regression of y onto $Lin(x, \mathbf{1})$ which results in $\hat{y}$. Following that, we project $\hat{y} = \hat{\beta}_1 + \hat{\beta}_2 x$ onto $Lin^{\perp}(\mathbf{1})$ which yields $\hat{\beta}_2 x^c$. After all, we translate $\sqrt{RSS}$ onto $Lin^{\perp}(\mathbf{1})$. These steps are demonstrated in Figure 34(a).

Looking at Figure 34(b) which depicts the $Lin^{\perp}(\mathbf{1})$, we derive

$$\text{ctg}\,\varphi = \frac{\hat{\beta}_2|x^c|}{\sqrt{RSS}} = t$$

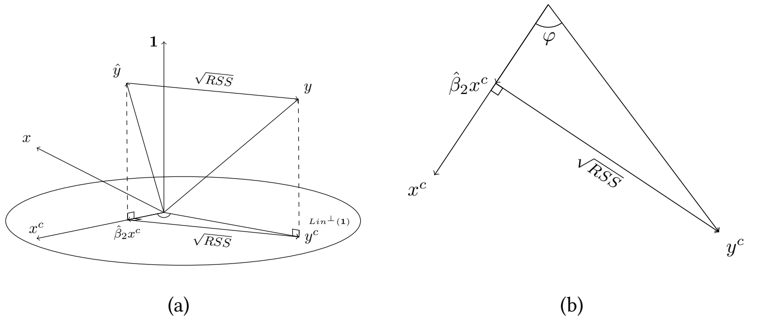


Figure 34: (a): Regression of y onto $Lin(x, \mathbf{1})$ and appropriate projections; (b): $Lin^{\perp}(\mathbf{1})$.

F-distribution

Generally, the following definition is given.

Definition 6. Let $\gamma_1 \sim \chi^2_{k_1}, \gamma_2 \sim \chi^2_{k_2}$, γ_1, γ_2 independent. Then

$$\frac{\gamma_1/k_1}{\gamma_2/k_2} \sim F_{k_1,k_2}.$$

Definition 7. Let $z = \begin{pmatrix} z_1 \\ \vdots \\ z_n \end{pmatrix}$ where $z_i \sim \mathcal{N}(0, \sigma^2)$ and are independent. Let L_1, L_2 be orthogonal subspaces in $\mathbb{R}^n$. then

$$F = \frac{\|H_1 z\|^2/dimL_1}{\|H_2 z\|^2/dimL_2} \sim F_{dimL_1,dimL_2},$$

where $\|H_1 z\|^2$, $\|H_2 z\|^2$ are the squared lengths of z projected onto L_1 and L_2 respectively.

Recall that by Theorem 12 the projections of a standard noraml vector onto orthogonal subspaces are independent. Thus, in terms of the Definition 7 $H_1 z$ and $H_2 z$ are independent. Next, from the Theorem 13 where we defined the chi-squared distribution it follows that the squared lengths of these projections follow the chi-squared distribution with the number of degrees of freedom equal to the dimension of the subspace onto which the vector was projected. In other words, $\|H_1 z\|^2 \sim \chi^2_{dimL_1}$, $\|H_2 z\|^2 \sim \chi^2_{dimL_2}$.

Taking the ratio of these length squared, we get the interpretaion of the angle between the original vector z and its projection onto L_1:

$$tg^2\varphi = \frac{\|H_1 z\|^2}{\|H_2 z\|^2}.$$

Adjusting this ratio to the degrees of freedom, we get the desired definition.

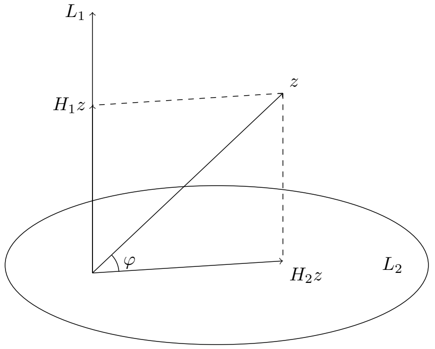


Figure 35: F-distribution as the ratio of the projection lengths squared adjusted to the dimensions of the subspaces.

F-test

The significance of several coefficients at once can be tested with the F-test. The F-statistic has the following form

$$F = \frac{(RSS_R - RSS_{UR})/q}{RSS_{UR}/(n - k_{UR})}$$

where indices R and UR stand for the restricted and unrestricted models respectively, n — number of observations, k — number of regressors, q — number of equtions used in the null hypothesis.

Due to plotting limitations, we consider the unrestricted model to be

$$y = \beta_1 + \beta_2 x + u$$

and the restricted model to be

$$y = \alpha_1 + v$$

Note that there was a choice in the restricted models.

We perform both regressions in order to get the ressiduals and plot them in Figure 36. Adjusted to the degrees of freedom, the ratio can be expressed in terms of the angle between two vectors, φ, as demonstrated in Figure 36

$$F = \frac{(RSS_R - RSS_{UR})/q}{RSS_{UR}/(n - k_{UR})} = \text{ctg}^2\,\varphi$$

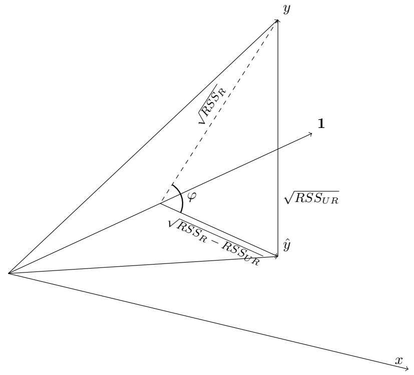


Figure 36: F-statistic as the cotangens squared of φ

ML

Principal Component Analysis

PCA is a tool to preserve as much variability in n-dimensional data as possible by projecting it onto q-dimensional subspace spanned by the principal components. There are two ways to find these principal components. One can either maximize the length of the projection of some vector x onto the q-dimensional subspace, or minimize the length of its orthogonal component. Both ways are equivalent which can be shown by the orthogonal decomposition theorem. However, there is also a geometrical interpretaion.

Theorem 14. *Let L be a k-dimensional subspace of $\mathbb{R}^n$. Then each vector x in $\mathbb{R}^n$ can be written uniqely in the form*

$$x = h + p$$

where p is in L and h is in $L^\perp$.

Proof. Let $e_1, \dots, e_k$ be an orthogonal basis for L, and define p as

$$p = \frac{x \cdot e_1}{e_1 \cdot e_1} + \ldots + \frac{x \cdot e_k}{e_k \cdot e_k}.$$

Being a linear combination of $e_1, \dots, e_k$, p is in L. Let $h = x - p$. Since each e_i, $i = 1, \dots, k$ is orthogonal to all the rest basis vectors, it follows that for all i

$$\begin{aligned} h \cdot e_i &= (x - p) \cdot e_i \\ &= x \cdot e_i - \left(\frac{x \cdot e_i}{e_i \cdot e_i}\right) \cdot e_i \underbrace{-0 - 0 \ldots - 0}_{(k-1)\text{ times}} \\ &= x \cdot e_i - x \cdot e_i = 0. \end{aligned}$$

Thus, we can conclude that h is in $L^\perp$.

To show that the decomposition stated in the theorem is unique, suppose there is another one: $x = h_1 + p_1$ where $h_1 \in L$ and $p_1 \in L$. Then $h + p = h_1 + p_1$ as both sides equal x and

$$p - p_1 = h - h_1.$$

It follow that the vector $v = p - p_1$ is in L and in $L^\perp$ which is possible if and only if $v = 0$. □

From this theorem it directly follows that

$$\|x\|^2 = \|h\|^2 + \|p\|^2.$$

Thus, holding the observation vector x fixed maximizing $\|h\|$ and minimizing $\|p\|$ are equivalent in order to get the longest projection of x onto L.

Theorem 15. *In PCA the best-fit q-dinesional subspace for a set of observations $x_1, \dots, x_n$ is defined by solving either*

$$\sum_{n=1}^{N} \|h\|^2 \to \min_{v_1, \dots, v_k}$$

which minimizes the distance between the observation and the subspace, or

$$\sum_{n=1}^{N} \|p\|^2 \to \max_{w_1, \dots, w_k}.$$

which maximizes the varince of projected data.

Proof. Consider a space of n observations and two factors. Additionaly assume that all the observations are centred and the best-fit line is already found.

Let us drop the perpendiculars from all the observation points onto the best-fit line and denote them as a_i, $i = 1, \dots, n$. Then there is a right triangle for each observation. We denote another leg as b_i and the distance of observation point from the origin as c_i. Thus, for every point we can apply the Pythagorean theorem: $c_i^2 = a_i^2 + b_i^2$. Summing through all the observations, we get

$$\sum_{i=1}^{n} c_i^2 = \sum_{i=1}^{n} a_i^2 + \sum_{i=1}^{n} b_i^2$$

Being unable to change the distance of the observation from the origin, we can either minimize $\sum_{i=1}^{n} a_i^2$, or maximize $\sum_{i=1}^{n} b_i^2$ in order to get the principal component.

□