How Gauss and Markov met Pythagoras: geometry in econometrics

Boris Demeshev[1], *Olga Gnilova*[2]

July 5, 2018

Contents

[1] boris.demeshev@gmail.com

[2] olyagnilova@gmail.com

Introduction

Pursuing abstractness and generality, a number of books and articles in econometrics rely heavily on standard algebraic proofs. However, most of the theorems proved in such a way have a strong geometric appeal. This paper demonstrates how shorter, less technical and even more beautiful the proofs can be when they are based on geometric theorems. We show that this technique can be extended to explain concepts in probability and statistics.

Although most of the theorems and ideas are not completely new and there are even a few works on the geometry in econometrics and statistics, this paper introduces alternative explanations and more general results in some cases. Further, there are algebraic proofs provided parallel to geometric ones and illustrations which are our own work. The illustrations are published at `https://github.com/olyagnilova/gauss-markov-pythagoras` and licensed under the Creative Commons Attribution 4.0. They are free to use and have the potential to serve as a pedagogical tool in explaining material for students.

Other researchers have done similar work, an especially thought-provoking and motivating paper was Cobb. The geometric proof of the Gauss-Markov theorem and the introduction of the Herschel-Maxwell approach are inspired by Cobb's work. Jacobson's (2011) thesis "The Geometry of the General Linear Model" is a comprehensive overview of a geometric approach in linear models including statistical foundations. Halliwell's (2015) text *The Gauss-Markov Theorem: Beyond the BLUE* and "An Intuitive Geometric Approach to the Gauss Markov Theorem" by Silva Pereira, Chaves, and Souza (2017) were helpful when thinking over alternative ways of proving this central theorem of econometrics. The works that deepened our understanding of instrumental variables included Butler's (2016) "The Simple Geometry of Correlated Regressors and IV Corrections" and lecture notes on Econometric theory 1 course by Ping Yu. The first work that proves the Frisch-Waugh-Lovell theorem using pure geometry is *A Geometric Representation of the Frisch-Waugh-Lovell Theorem* by Sosa Escudero (2001). Finally, there were several works that apply the geometric approach to hypothesis testing. They are "The geometrical interpretation of statistical tests in multivariate linear regression" by Langsrud (2004), "On the geometry of F, Wald, LR, and LM tests in linear regression models" by Siniksaran (2005), and one which uses an unusual approach — "Elliptical Insights: Understanding Statistical Methods through Elliptical Geometry" by Friendly, Monette, and Fox (2013).

The paper has four parts. In the first the fundamentls of the geometry of random variables are introduced with examples. In part two, we develop the ideas related to linear regressions from the ideas which need almost no assumptions to more sophisticated theorems and concepts. The third part is devoted to partial correlations. It contains a proof of a newly-introduced fact about the partial correlation and the correlation between the residuals in alinear regression model. Finally, probability distributions are introduced from a geometric perspective and hypothesis tests are illustrated.

Geometric interpretation of random variables

It is fundamental to define the geometric properties of random variables using 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) as 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 not difficult to check that the definition satisfies the properties of a scalar product assuming that X and Y are the same random variables if there is a constant a such that $\mathbb{P}(X = Y + a) = 1$.

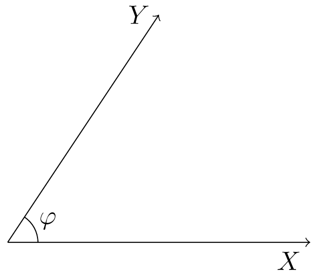


Figure 1: Geometric representation of random variables.

Having defined the scalar 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 standard deviation of X (σ_X) is the length.

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 iff their scalar product equals 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 $\cos(a, b)$ 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 Pythagorean 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:

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

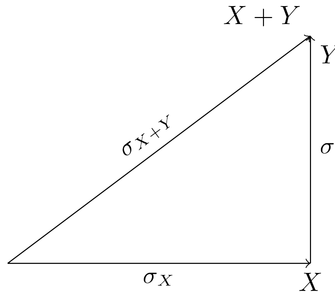


Figure 2: The Pythagorean theorem for random variables X and Y.

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

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

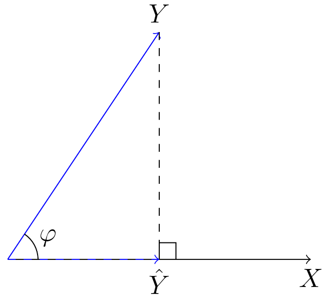


Figure 3: The projection of a random variable Y onto the line spanned by a random variable X.

The law of iterated expectations

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

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

Proof. Consider a vector space of all the random variables. The random variables which can be described as functions 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 of all the functions of X. As a result of this step, we get $\text{E}(Y|X)$ — the function of X that predicts Y the best (the function which gives the lowest MSE). Next, projecting $\text{E}(Y|X)$ onto the space of all constants, we obtain $E(Y)$.

Notice that the vector $Y - \text{E}(Y|X)$ (which is also called the residual) is perpendicular to the plane α. Particularly, 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.

MSE decomposition

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

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

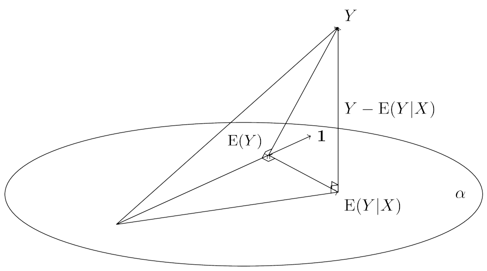


Figure 4: The law of iterated expectations. Equivalence of the two-step projecttion and direct projection of Y onto $\mathbf{1}$.

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:

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

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

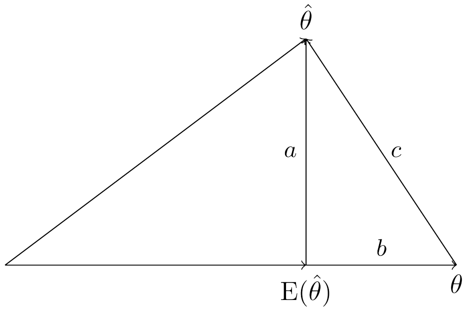


Figure 5: Decomposition of mean squred error into the variance and the bias squared $\left(a = \sqrt{\|\hat{\theta} - \mathrm{E}(\hat{\theta})\|^2},\right.$ $\left. b = \sqrt{\|\theta - \mathrm{E}(\hat{\theta})\|^2},\ c = \sqrt{\|\hat{\theta} - \theta\|^2}\right)$

□

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 it was done in the previous 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:

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

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

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 $\hat{y} = \text{sCorr}(x, y) \cdot y$.

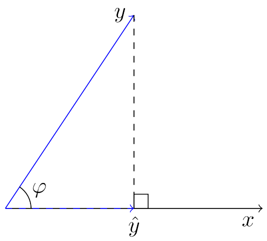


Figure 6: Vector y projected onto vector x.

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

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

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

$$\begin{aligned} \text{sCorr}(x + \alpha\mathbf{1}, y) &= \frac{\text{sCov}(x + \alpha\mathbf{1}, y)}{\sqrt{\text{sVar}(x + \alpha\mathbf{1})\,\text{sVar}(y)}} \\ &= \frac{\text{sCov}(x, y) + \text{sCov}(\alpha\mathbf{1}, y)}{\sqrt{\text{sVar}(x)\,\text{sVar}(y)}} \\ &= \frac{\text{sCov}(x, y)}{\sqrt{\text{sVar}(x)\,\text{sVar}(y)}} \\ &= \text{sCorr}(x, y) \end{aligned}$$

Proof. Firstly, we project vectors x and y onto $\text{Lin}^{\perp}(\mathbf{1})$ 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

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

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.

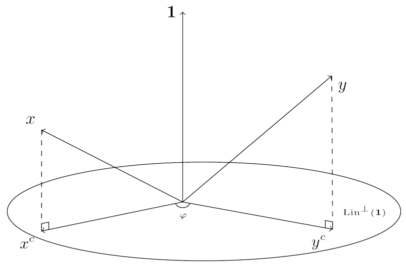


Figure 7: Centred vectors x^c and y^c.

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 $\text{Lin}^{\perp}(\mathbf{1})$. 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 $\text{Lin}^{\perp}(\mathbf{1})$ yields zero as demonstrated in Figure . Moreover, it follows that the angle between $\tilde{x}$ and y is still φ.

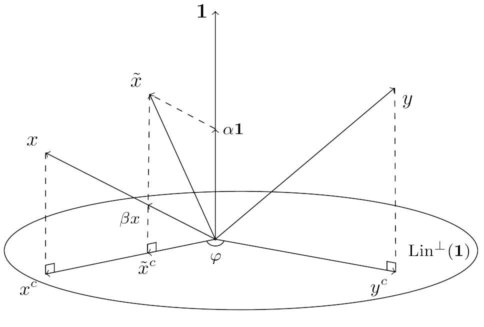


Figure 8: $\text{sCorr}(x + \alpha\mathbf{1}, y) = \text{sCorr}(x, y)$ as the corresponding angles are equal.

Finally, putting everything together we finish the proof:

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

□

Sample correlation coefficient in simple linear regression

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

$$y = \beta_1 \mathbf{1} + \beta_2 x + \varepsilon$$

has the property

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

Proof. Firstly, we consider the case when $\hat{\beta}_2 > 0$. 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 plane perpendicular to the vector of all ones (denoted as $\mathbf{1}$) as shown in Figure 9(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 $\vec{a}$ onto the span of a vector of all ones yields the vector of averages $\bar{\vec{a}}$.

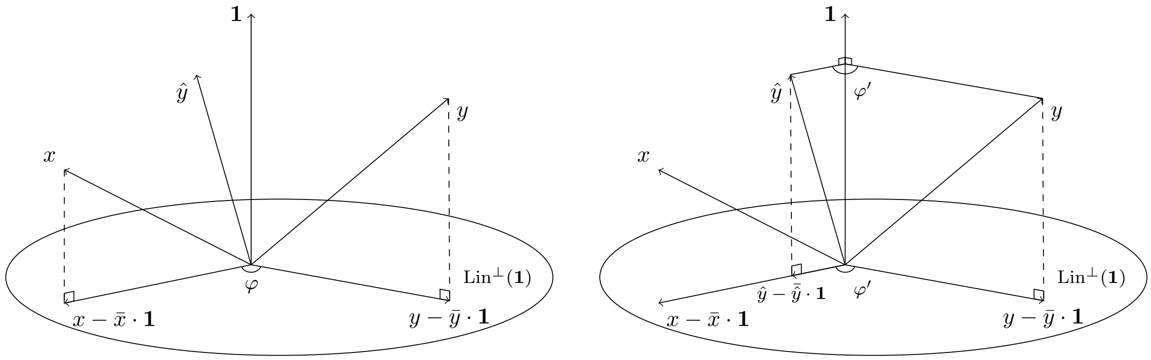
In order to get the angle between y and $\hat{y}$ we should start with regressing y on $\text{Lin}(x, \mathbf{1})$. Then the only thing left is to project $\hat{y}$ onto $\text{Lin}^{\perp}(\mathbf{1})$ since the y vector has already been projected. Note that the projected $\hat{y}$ falls onto tha span of vector $x - \bar{x} \cdot \mathbf{1}$ as it can be decomposed into a sum $ax + b\mathbf{1}$ where $a, b \in \mathbb{R}$. The first component 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 9(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

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


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

(a)

(b)

Figure 9: (a): 'Centred' x and y, i.e., projected onto $\text{Lin}^{\perp}(\mathbf{1})$; (b): 'Centred' $\hat{y}$, i.e., projected onto $\text{Lin}^{\perp}(\mathbf{1})$.

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 $\text{sCorr}(x, y) = \text{sCorr}(y, \hat{y})$.

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 between 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$(y, \hat{y})$.

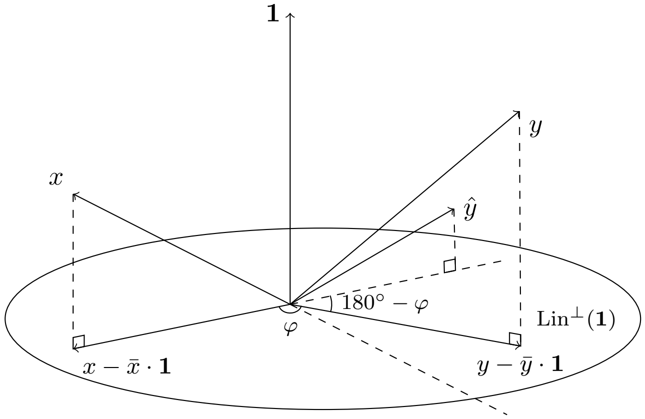

$\mathbf{1}$
y
x
$\hat{y}$
$180° - \varphi$
φ
$\text{Lin}^{\perp}(\mathbf{1})$
$x - \bar{x} \cdot \mathbf{1}$
$y - \bar{y} \cdot \mathbf{1}$

Figure 10: 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 11(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 regression 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

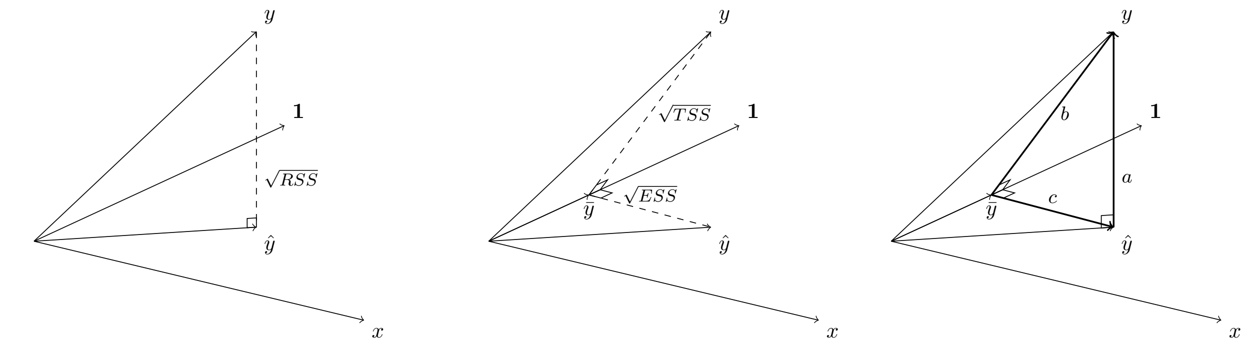
$$\hat{\beta} = (X^T X)^{-1} X^T y$$

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:

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



(a) (b) (c)

Figure 11: (a): Residual sum of squares; (b): Total sum of squares and residual sum of squares; (c): Illustration of the equality $(\sqrt{RSS})^2 + (\sqrt{ESS})^2 = (\sqrt{TSS})^2$ where a stands for $\sqrt{RSS}$, $b - \sqrt{TSS}$, $c - \sqrt{ESS}$.

The final step is to put everything together. Note that since $y - \hat{y}$ is perpendicular to $\text{Lin}(x, \mathbf{1})$ it is also perpendicular to $\hat{y} - \bar{y}$ and $\mathbf{1}$ as these vectoros are in $\text{Lin}(x, \mathbf{1})$. 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 $\sqrt{TSS}$:

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

□

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

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

When putting everything together all the terms cancel out which proves

$$ESS + RSS = TSS$$

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

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

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 $\hat{y}$.

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

Consider Figure 11(c) 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 12.

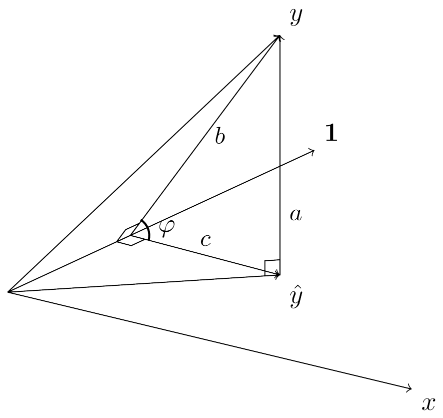


Figure 12: Determination coefficient as squared $\cos \varphi$ where a stands for $\sqrt{RSS}$, b – $\sqrt{TSS}$, c – $\sqrt{ESS}$.

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 $\cos \varphi$. 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.

The first step is regressing y on $\text{Lin}(\mathbf{1}, x)$. As shown in Figure 13(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 13(b).

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

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

Premultiplying both sides by X^T, we obtain:

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

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

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

we obtain

$$\sum_{i=1}^{n} \hat{y}_i = \sum_{i=1}^{n} y$$

And this finishes the proof:

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

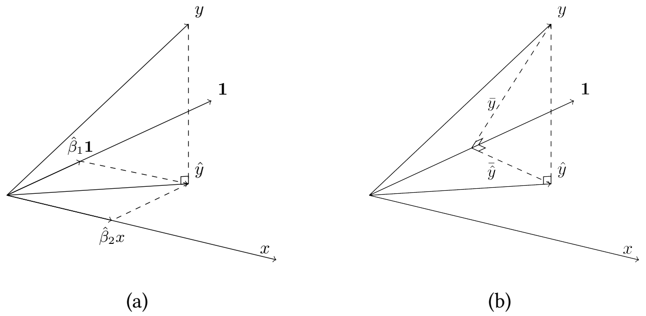


Figure 13: (a): Regression of y on $\text{Lin}(\mathbf{1}, x)$; (b): Regression of y and $\hat{y}$ on $\mathbf{1}$.

□

Orthogonality of regressors

Theorem 8. *Omitting a variable in a regression model does not lead to a bias in coefficient estimators if either the coefficient of the omitted regressor equals zero or the regressors are uncorrelated.*

Proof. We will not consider the case when the coefficient of the omitted regressor equals zero because this means that the model is not mis-specified. The case of orthogonal regressors is less trivial. However, it can be easily proved once the geometric approach is implemented.

Condider the following regression model

$$y = \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

where regressors x_1 and x_2 are uncorrelated, $x_1 \perp x_2$. When we perform a regression of y onto both x_1, x_2 we obtain predicted values $\hat{y}$ which can be broken up into a sum:

$$\hat{y} = \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2.$$

Since x_1 and x_2 are orthogonal it folows that the difference $\hat{y} - \hat{\beta}_1 x_1$ must be orthogonal to x_1. Thus, by the theorem of three perpendiculars the projection of y onto x_1 also results in $\hat{\beta}_1 x_1$. □

Consider a model

$$y = \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

with centred regressors x_1, x_2 such that $\bar{x}_1 = \bar{x}_2 = 0$. Suppose we mistakenly believe in a model

$$y = \beta_1^* x_1 + \varepsilon^*$$

where x_2 is omitted. Then the new estimator

$$\begin{aligned} \mathrm{E}\left(\hat{\beta}_1^*\right) &= \mathrm{E}\left(\frac{\mathrm{sCov}(x_1, y)}{\mathrm{sVar}(x_1)}\right) \\ &= \mathrm{E}\left(\frac{\mathrm{sCov}(x_1, \beta_1 x_1 + \beta_2 x_2 + \varepsilon)}{\mathrm{sVar}(x_1)}\right) \\ &= \beta_1 + \beta_2 \frac{\mathrm{sCov}(x_1, x_2)}{\mathrm{sVar}(x_1)} + \mathrm{E}\left(\frac{\mathrm{sCov}(x_1, \varepsilon)}{\mathrm{sVar}(x_1)}\right) \\ &= \beta_1 + \beta_2 \frac{\mathrm{sCov}(x_1, x_2)}{\mathrm{sVar}(x_1)} \end{aligned}$$

will be unbiased if either $\beta_2 = 0$ or $\mathrm{sCov}(x1, x2) = 0$.

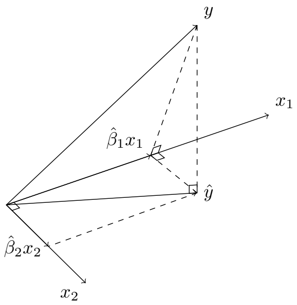


Figure 14: In case of uncorrelated regressors omitting one of them does not result in bias of the estimator.

Geometry of formula for OLS estimators

The following interpretation gives another view on the estimators obtained by the ordianry least squares method. Previously, we illustrated the ideas in the observations space, $\mathbb{R}^n$. Apparantely, the OLS estimators can be depicted in the regressors space, $\mathbb{R}^k$. This result follows from the Cramer's rule.

Consider a linear model with two regressors:

$$\hat{y} = \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2.$$

Let us denote $X = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$ and $\langle X, y \rangle = \begin{pmatrix} \langle x_1, y \rangle \\ \langle x_2, y \rangle \end{pmatrix}$. Then, taking scalar product, we obtain:

$$\langle X, y \rangle = \hat{\beta}_1 \langle X, x_1 \rangle + \hat{\beta}_2 \langle X, x_2 \rangle$$

All the elements of this equation are illustrated in Figure .

The area of each parallelogram is the determinant of the matrix with columns which form the corresponding parallelogram. For instance, the area of parallelogram formed by vectors $\langle X, y\rangle$, $\langle X, x_2\rangle$ is equal to $\det([\langle X, y\rangle\langle X, x_2\rangle])$. The parallelogram formed by vectors $\langle X, \hat{\beta}_1 x_1\rangle$, $\langle X, x_2\rangle$ has the area of $\hat{\beta}_1 \det([\langle X, x_1\rangle\langle X, x_2\rangle])$. Since both parallelograms must be of the same are we obtsin a formula for $\hat{\beta}_1$:

$$\hat{\beta}_1 = \frac{\det([\langle X, y\rangle\langle X, x_2\rangle])}{\det([\langle X, x_1\rangle\langle X, x_2\rangle])}.$$

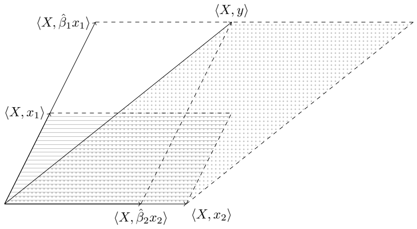

$\langle X, \hat{\beta}_1 x_1\rangle$
$\langle X, y\rangle$
$\langle X, x_1\rangle$
$\langle X, \hat{\beta}_2 x_2\rangle$
$\langle X, x_2\rangle$

Figure 15: Illustration of estimators in $\mathbb{R}^k$ as the ratio of parallelogram areas.

So as to generalize this result we introduce a matrix $A = (y|X)$ that has y as its first column and all the regressors as the rest ones. We will denote this matrix without column y or x as A_{-y}, A_{-x}, respectively. Then, we can write out the formula for β_x:

$$\hat{\beta}_x = \frac{\det(\langle A_{-y}, A_{-x}\rangle)}{\det(\langle A_{-y}, A_{-y}\rangle)}.$$

If there is a regression with unit vector among the regressors and the Frisch-Waugh-Lovell theorem is applied, the formula above takes the followin form:

$$\hat{\beta}_x = \frac{\det(\text{sCov}(A_{-y}, A_{-x}))}{\det(\text{sVar}(A_{-y}))}.$$

Frisch–Waugh–Lovell theorem

Theorem 9. *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 comformable, $k_1 \times 1$ and $k_2 \times 1$ vectors. Consider another regression

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

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).

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 a 'one-step' regression and will distinct its coefficients with an upper index A. The only step in obtaining β_1^A is regressing y

From regression (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 regression (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

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

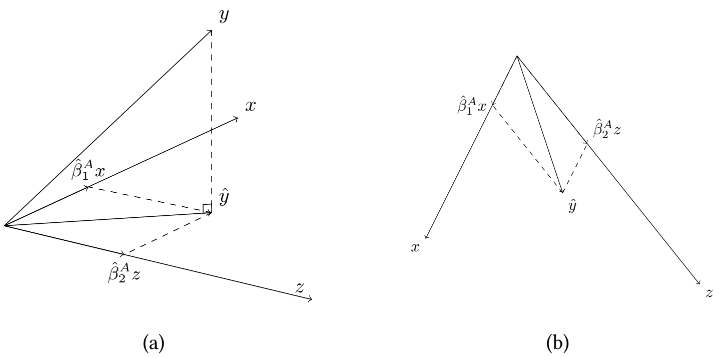


Figure 16: (a): Regression of y on Lin(x, z); (b): Lin(x, z).

on Lin(x, z) and then expanding $\hat{y}$ as a linear combination of basis vectors x and z, which is shown in Figure 16(a). Figure 16(b) depicts 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. We will denote its coefficients with an upper index B.

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

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 17(a).

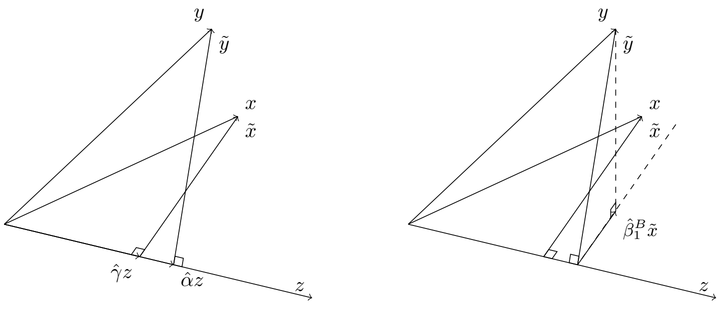
Finally, 'cleansed' y must be regressed on 'cleansed' x. However, it cannot be performed immediately as $\tilde{y}$ and $\tilde{x}$ are skew lines. So at first, we fix this problem by translation and after that obtain $\hat{\beta}_1^B \tilde{x}$ (see Figure 17(b)).

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

In Figure 18(b) segments AF and $BH = DG$ stand for $\hat{\beta}_1^A x$ and $\hat{\beta}_1^B \tilde{x}$ respectively, while segments AC and BC represent x and $\tilde{x}$. Having two congruent angles, triangles ABC and FHC are simillar. Then, it follows:

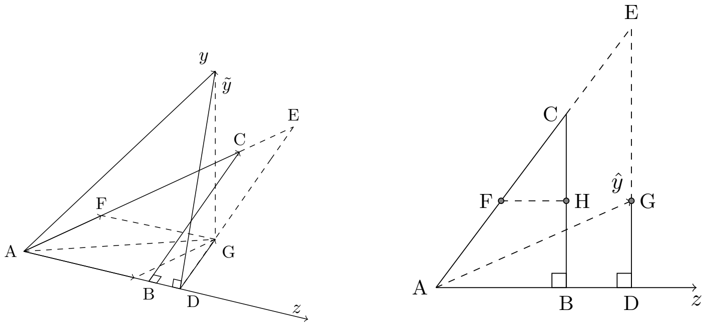
$$\frac{AF}{AC} = \frac{BH}{BC} \Leftrightarrow \frac{\hat{\beta}_1^A x}{x} = \frac{\hat{\beta}_1^B \tilde{x}}{\tilde{x}} \Leftrightarrow \hat{\beta}_1^A = \hat{\beta}_1^B$$

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



(a) (b)

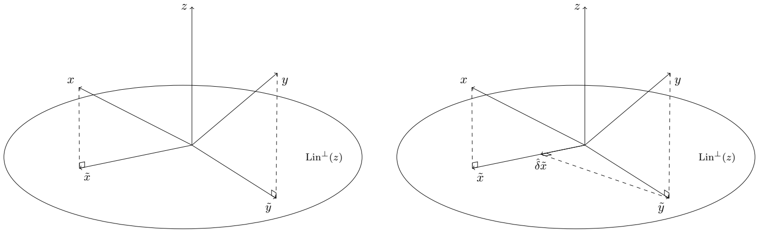
Figure 17: (a): Regression of y on z and of x on z; (b): Translation of $\tilde{x}$.



(a) (b)

Figure 18: (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 — $\hat{\beta}_1^A x$, G — $\hat{\beta}_1^B \tilde{x}$; (b): Lin$(x, z)$.

orthogonal to z, i.e., $\text{Lin}^{\perp}(z)$ as demonstrated in Figure 19(a). Then we perform a regression of the 'cleansed' $\tilde{y}$ on the 'cleansed' $\tilde{x}$ (see Figure19(b)).



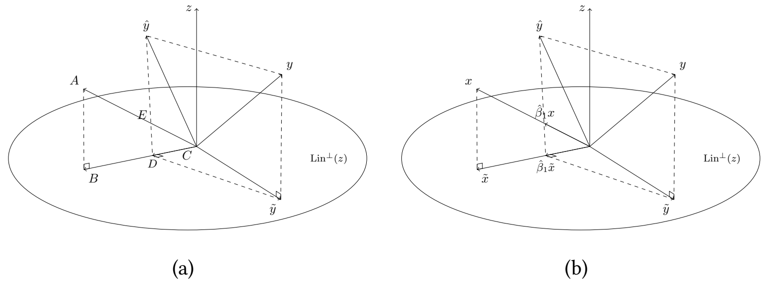
(a) (b)

Figure 19: (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 exactly 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 20: (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 to 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 10. *Let x_i be a $n \times 1$ regressor, u_i — a residual in regression of x_i on all the rest regressors, $i = 1, \dots, k$. Consider a transformation of a vector v, $f(v) = v/\|v\|^2$. Then applying this transformation on the residuals $u_1, \dots, u_k$ yields new regressors $v_1, \dots, v_k$. 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 $x_1, \dots, x_k$.*

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

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

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

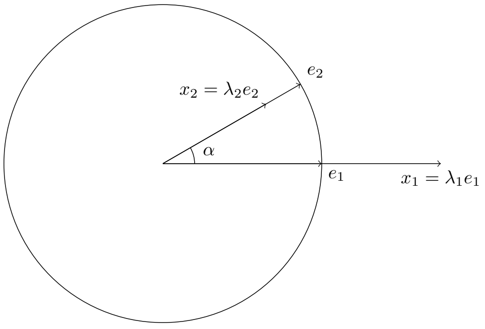


Figure 21: Two regressors in the unit circle.

Then we perform two regressions

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

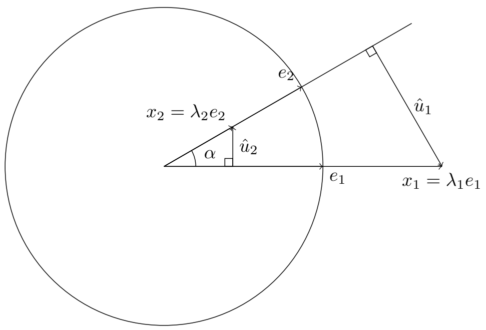


Figure 22: Residuals $\hat{u}_1$ and $\hat{u}_2$.

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

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

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

For convenience 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 24(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

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

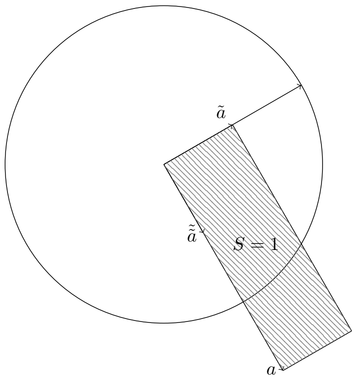


Figure 23: Example of inversion for vector a.

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.

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

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

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

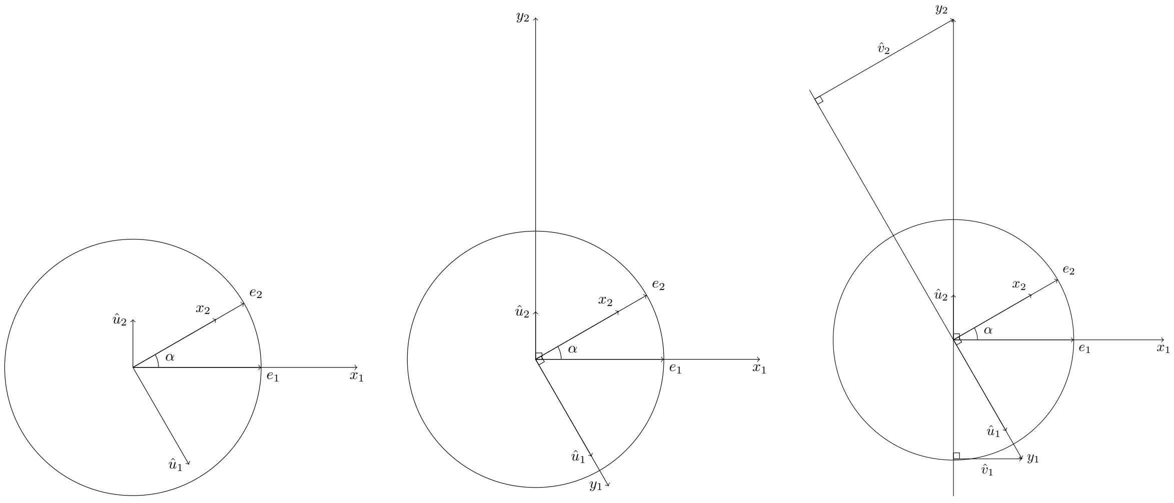
Next, we perform another two regressions:

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

There are 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 24(c):

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



(a) (b) (c)

Figure 24: (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.

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:

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

□

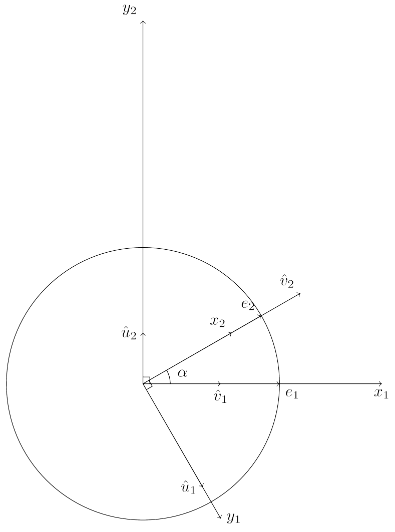


Figure 25: New residuals translated to the origin of the unit circle.

Gauss-Markov theorem

Theorem 11. *In the homoskedastic linear regression model the best (minimum-variance) linear unbiased estimator is given by the ordinary least squares.*

Proof. Consider an OLS estimator and an alternative one:

$$\hat{\beta}_{OLS} = (X^T X)^{-1} X^T y = A^T y$$
$$\hat{\beta}_{alt} = A^T_{alt} y$$

Note that $A^T X = I_k$, then the following holds for all β:

$$A^T X\beta = \beta$$
$$A^T_{alt} X\beta = \beta$$

Taking the difference of these equations, we obtain:

$$\left(A^T_{alt} - A^T\right) X\beta = 0 \Rightarrow \left(A^T_{alt} - A^T\right) \perp X$$

If we treat the coefficients separately and consider, for instance, $\beta^{(2)}$, we get the following result

$$\left(a^T_{alt} - a^T\right) \perp X$$

where a_{alt} and a are the second columns of matrices A_{alt} and A correspondingly. Since $a \in \text{Lin}(\text{col}\, X)$, it follows that $a_{alt} \notin \text{Lin}(\text{col}\, X)$.

Now we can express the variances of both estimators in terms of a_{alt} and a:

$$\text{Var}\left(\hat{\beta}^{(2)}_{OLS}\right) = \text{Var}\left(a^T y\right) = a^T \sigma^2 I_k a = \sigma^2 \left\|a\right\|^2$$
$$\text{Var}\left(\hat{\beta}^{(2)}_{alt}\right) = \text{Var}\left(a^T_{alt} y\right) = a^T_{alt} \sigma^2 I_k a_{alt} = \sigma^2 \left\|a_{alt}\right\|^2$$

Since vectors a, a_{alt} and $a - a_{alt}$ form a right triangle and $a_{alt} \notin \text{Lin}(\text{col}\, X)$, the vector $\left\|a_{alt}\right\|^2$ must be longer than a, and the corresponding estimator must have higher variance.

□

Consider an estimator β which is a linear function of Y:

$$\hat{\beta} = A^T Y$$

where A is an $n \times k$ function of X such that $A^T X = I_k$. From

$$\text{Var}(\hat{\beta}_{OLS}) = (X^T X)^{-1} \sigma^2$$
$$\text{Var}(A^T y) = A^T A \sigma^2$$

it follows that it is sufficient to prove that $A^T A - (X^T X)^{-1}$ is positive semi-definite. Set $C = A - X(X^T X)^{-1}$ and note that $X^T C = 0$, then

$$A^T A - (X^T X)^{-1}$$
$$= (C + X(X^T X)^{-1})^T (C + X(X^T X)^{-1}) - (X^T X)^{-1}$$
$$= C^T C + C^T X(X^T X)^{-1} + (X^T X)^{-1} X^T C +$$
$$(X^T X)^{-1} X^T X (X^T X)^{-1} - (X^T X)^{-1}$$
$$= C^T C$$

Matrix $C^T C$ is positive semi-definite since

$$\forall a \neq 0 \qquad a^T C^T C a = \|Ca\|^2 \geq 0$$

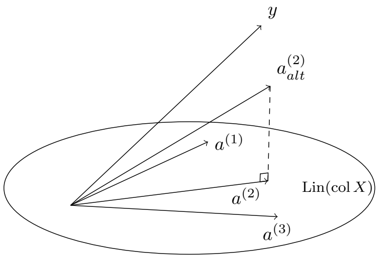


Figure 26: Gauss-Markov theorem for the case of three regressors where $a^{(1)}$, $a^{(2)}$, $a^{(3)}$ are columns of matrix A.

Geometry of instrumental variables

Consifder a model with an endogenity problem, i.e. explanotary variable x is correlated with the error term u:

$$y = \beta x + u$$

Assume there is an instrument z which is dependent with the problematic regressor x but uncorrelated with the error term u. The 2SLS procedure tells us to perform the following steps.

1. Regress x onto z and get the vector of predicted values $\hat{x}$,

2. Regress y onto $\hat{x}$.

These steps result in β_{IV} estimator which is illustrated in Figure 27.

The same result could be obtained with the oblique projection. That is projecting y onto x along the vector perpendicular to the span of z.

The equivalence of these two methods holds due to the similarity of triangles. Consider a plane α which satisfies the property of being perpendicular to the span of z, $\alpha \perp z$. Vectors x, z and $\hat{x}$ form a triangle which is denoted as $\triangle OBC$ where $\overrightarrow{OB} = \hat{x}$, $\overrightarrow{OC} = x$. In order to get an oblique projection of y onto x we could either project y directly onto x staying in the plane α or get the same result in two steps. First, project y onto z and then project the result onto x which gives the same outcome by the theorem of three perpendiculars. Thus, we get another triangle $\triangle OAD$ where $\overrightarrow{OA} = \hat{\beta}_{IV}\hat{x}$, $\overrightarrow{OD} = \hat{\gamma}x$. Since triangles $\triangle OBC$ and $\triangle OAD$ are similar it follows that

$$\frac{OD}{OC} = \frac{OA}{OB}$$

which means that

$$\hat{\gamma} = \hat{\beta}_{IV}.$$

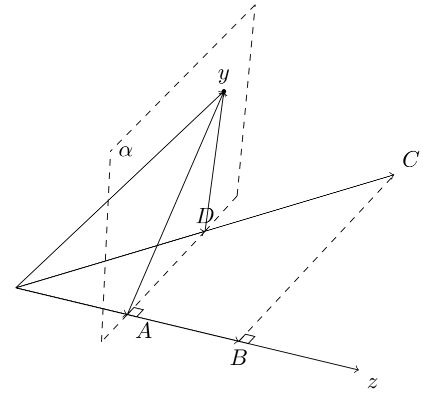


Figure 27: Geometry of instrumental variables. A stands for $\hat{\beta}_{IV}\hat{x}$, $B - \hat{x}$, $C - x$, $D - \hat{\beta}_{IV}x$.

Geometry of proxy variables

Consider a model

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

where the error term u is not correlated with the regressors. Suppose that w is an unobservable variable. One way to deal with this problem and get a consistent estimator of β_1 is to use a proxy variable. In order to clearly state its properties we decompose the unobservable variable into a sum of a multiple of the proxy (pr) and a part that is uncorrelated with the proxy ($\hat{\nu}$):

$$\hat{w} = \gamma \cdot pr + \hat{\nu}$$

Then the proxy variable must satisfy the following properties.

1. It must be correlated with the unobservable variable w, $pr \not\perp w$.

2. The error term $\hat{u}$ must be uncorrelated with the proxy variable, $pr \perp \hat{u}$.

3. The error term $\hat{\nu}$ must be uncorrelated with the regressor x, $x \perp \hat{\nu}$.

To get a consistent estimator of β_1 we need to regress y onto x and pr. Consistency is illustarted in Figure 28. Suppose we could get the $\hat{y}$ by performing the original regression of y onto x and w. Then, $\hat{y}$ could be decomposed in a sum of $\hat{\beta}_1 x$ and $\hat{\beta}_2 w$. Notice, that $\hat{y}$ is both projection of y onto w, x and onto w, x, pr due to the property of $\hat{\nu}$ being orthogonal to x. When projected onto the plane spanned by x and pr, the $\hat{\beta}_1 x$ component stays the same as it is already in this plane while $\hat{\beta}_2 w$ projects onto the span of pr by the second property of proxy variable.

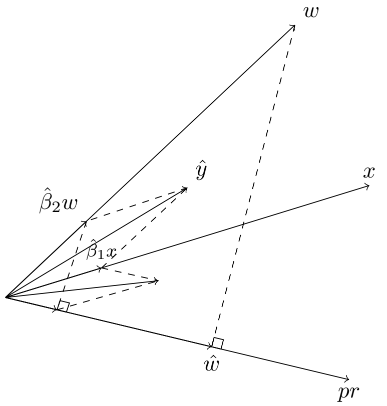


Figure 28: Geometry of proxy variables.

Partial correlations

Definition of a partial correlation

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

A partial correlation is a measure of the 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. A 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 a 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 $\text{Lin}^{\perp}(Z)$. We will call this variable 'cleansed' and label it as $\widetilde{X}$. Applying the same procedure for Y yields the 'cleansed' variable $\tilde{Y} = Y - \text{E}(Y|Z) \in \text{Lin}^{\perp}(Z)$. The angle between $\widetilde{X}$ and $\widetilde{Y}$ (φ in Figure) is the correlation coefficient between these 'cleansed' random variables and the partial correlation between the original ones.

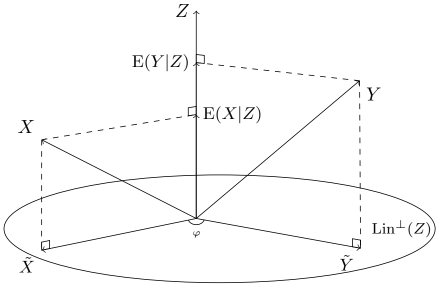


Figure 29: Partial correlation between X and Y while Z is fixed.

Definition 2. A partial correlation 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

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

The 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 $\hat{X} = \text{E}(X|Y, Z)$ can be broken up into the sum of $\beta_{XY} Y$ and $\beta_{XZ} Z$. Projecting $\beta_{XY} Y$ onto $\text{Lin}^{\perp}(Z)$ results in a vector $\alpha_{XY} \widetilde{Y}$ where $\widetilde{Y} = Y - \text{E}(Y|Z)$ is the projection of Y onto $\text{Lin}^{\perp}(Z)$.

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:

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

In the same manner we get α_{YX}:

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

Note, that α_{XY} and α_{XY} are of the same sign.

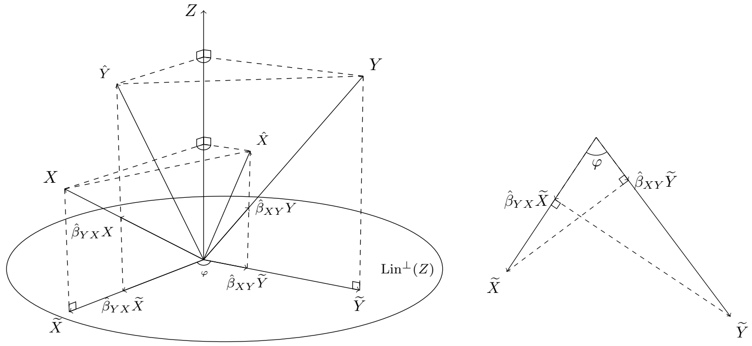
Multiplying these coefficients, we get the final result:

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

By the properties of similar triangles

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

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:


Z
$\hat{Y}$
Y
$\hat{X}$
X
$\hat{\beta}_{XY}Y$
$\hat{\beta}_{YX}X$
$\mathrm{Lin}^{\perp}(Z)$
φ
$\hat{\beta}_{XY}\widetilde{Y}$
$\beta_{YX}\widetilde{X}$
$\widetilde{Y}$
$\widetilde{X}$
φ
$\hat{\beta}_{XY}\widetilde{Y}$
$\hat{\beta}_{YX}\widetilde{X}$
$\widetilde{X}$
$\widetilde{Y}$

(a) (b)

Figure 30: (a): Alternative definition of the partial correlation; (b): $\mathrm{Lin}^{\perp}(Z)$.

Having plotted $\mathrm{Lin}^{\perp}(Z)$, now we can express $\cos\varphi$ in terms of β_{XY} and β_{YX}. We use the fact that $\beta_{XY}\beta_{YX} > 0$:

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

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

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

Partial correlation as correlation between residuals

Theorem 12. *The partial correlation between X and Y holding Z fixed is the negative correlation coefficient between the residuals u in the regression model*

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

and the residuals v in the model

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

Let us define cleansed X and Y first as

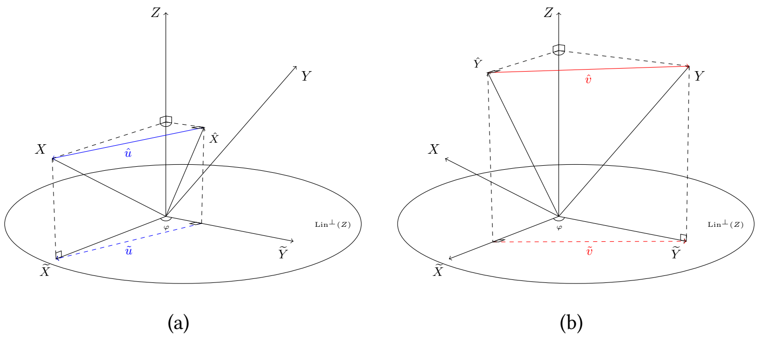
$$\widetilde{X} = \alpha_1\widetilde{Y} + \tilde{u}, \widetilde{X} \perp Z$$
$$\widetilde{Y} = \beta_1\widetilde{X} + \tilde{v}, \widetilde{Y} \perp Z$$

Then

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

Proof. The first step is to find the residuals in the regressions. For example, in order to get u we regress X onto $\mathrm{Lin}(Y, Z)$ which results in $\hat{X} = X - \mathrm{E}(X|Y, Z)$. Then we take the difference $X - \hat{X} = u$ and project it and X itself onto $\mathrm{Lin}^{\perp}(Z)$ as demonstrated in Figure 31(a). We denote the result as $\tilde{u}$ and $\widetilde{X}$ respectively.

Figure 31(b) shows the same steps for obtaining $\tilde{v}$ and $\widetilde{Y}$.



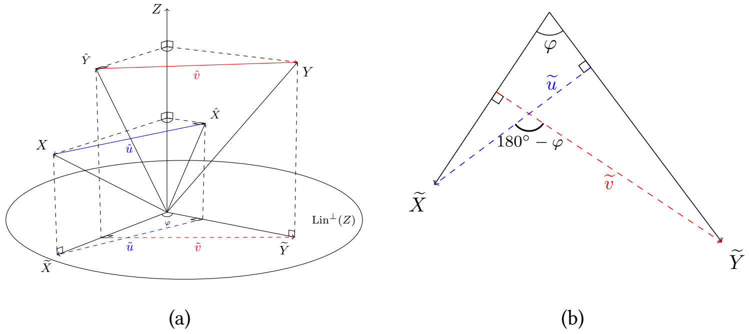
(a)

(b)

Figure 31: (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 32(b), we conclude that the desired angle is the bigger of the vertical angles. Hence, we can derive it from the property of the quadrilateral by substracting all the known angles from 360°. Thus, the desired angle is $180° - \varphi$.

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



(a)

(b)

Figure 32: (a): The residuals of both regressions; (b): $\mathrm{Lin}^{\perp}(Z)$.

□

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

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

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

$$\begin{aligned} \mathrm{Var}(\tilde{u}) &= \mathrm{Var}(\widetilde{X}) - \frac{\mathrm{Cov}^2(\widetilde{X}, \widetilde{Y})}{\mathrm{Var}^2(\widetilde{Y})}\,\mathrm{Var}(\widetilde{Y}) \\ &\quad - 2\,\mathrm{Cov}(\widetilde{X}, \widetilde{Y})\frac{\mathrm{Cov}(\widetilde{X}, \widetilde{Y})}{\mathrm{Var}(\widetilde{Y})} \\ &= \mathrm{Var}(\widetilde{X})\left(1 - \frac{\mathrm{Cov}^2(\widetilde{X}, \widetilde{Y}}{\mathrm{Var}(\widetilde{X})\,\mathrm{Var}(\widetilde{Y})}\right) \\ \mathrm{Var}(\tilde{v}) &= \mathrm{Var}(\widetilde{Y})\left(1 - \frac{\mathrm{Cov}^2(\widetilde{X}, \widetilde{Y}}{\mathrm{Var}(\widetilde{X})\,\mathrm{Var}(\widetilde{Y})}\right) \end{aligned}$$

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

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

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.

The original idea belongs to J.C. Maxwell who was wondering what distribution the velocity of gas molecules follow. The similar question was also bothering J.H.W. Herschel who was an astronomer and was dealing with measurement errors in astronomical data. Here we provide a proof of the theorem known today as Herschel-Maxwell's.

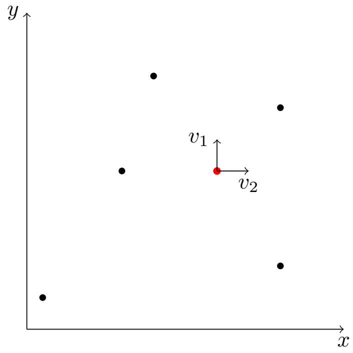


Figure 33: Black dots represent the gas molecules. The red dot stands for the one we catch. Its speed along the horizontal axis is v_1, i.e., the first component of the velocity vector, and its speed along the vertical axis is v_2.

Assume there are gas molecules moving chaotically on a plane and we can measure the velocity vector of one of them. We will denote this vector as $V = \begin{pmatrix} X \\ Y \end{pmatrix}$ where X and Y stand 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;
3. And we measure the velocity in such a way that $\text{Var}(X) = 1$.

Theorem 13. *Assumptions (1), (2) and (3) are satisfied if and only if* $X \sim \mathcal{N}(0, 1)$*,* $Y \sim \mathcal{N}(0, 1)$ *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 $X \sim Y$. It follows that

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

which holds for $\text{E}(X) = \text{E}(Y) = 0$ only. One may notice that $\text{E}(X)$ may not exist at all, but we derive later the exact distribution of X.

Likewise, $\text{Var}(X) = \text{Var}(Y)$ and it follows from the assumption (3) that $\text{Var}(X) = \text{Var}(Y) = 1$.

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 variable U is uniform on the interval $(0, 2\pi)$.

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

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.

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 order 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 $f(x, y) = ce^{-\frac{x^2+y^2}{2}}$ 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:

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

Without assumption (3) variables X and Y are normal $\mathcal{N}(0, \sigma^2)$. □

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

1. *the function* $f(z)$ *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

Definition 6. 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 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.

Theorem 15. *The definitions 5 and 6 are equivalent.*

Proof. First, it can be shown that the projected vector $\hat{z}$ is the original vector z multiplied by the projection matrix $H = X(X^TX)^{-1}X^T$ where the columns of X are fixed linearly independent vectors $x_1, \ldots, x_k$ in L or equivalently col X = 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 $D = \text{diag}\,(\lambda_1, \ldots, \lambda_n)$ 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 col X. Then for any x_i, $i = 1, \ldots, k$, $Hx_i = x_i \cdot 1$ since any x_i is already in col X. This implies that $\lambda_1 = \ldots = \lambda_k = 1$. There are also $n - k$ vectors in the subspace orthogonal to col X. So for any x_i, $i = k + 1, \ldots, n$, the orthogonal projection yields zero. We conclude that $\lambda_{k+1} = \ldots = \lambda_n = 0$.

Rewritting the theorem statement further, we obtain

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

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

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

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

□

Definition 5. 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

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

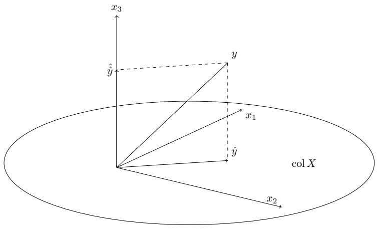


Figure 34: Consider a 3-dimensional example, col X = Lin(x_1, x_2) and $col^{\perp} X$ = Lin(x_3). $Hx_1 = x_1$ and $Hx_2 = x_2$ since they are in col X. However, $Hx_3 = 0$ as $x_3 \perp$ col X. Projecting an arbitrary vector onto col X yileds $Hy = \hat{y} \in$ Lin(x_1, x_2) while projecting onto col$^{\perp}$ X results in $(I - H)y = \hat{\hat{y}} \in$ Lin(x_3).

Student's

Definition 7. Let $z = \begin{pmatrix} z_1 \\ \vdots \\ z_n \end{pmatrix}$ where $z_i \sim \mathcal{N}(0, \sigma^2), i = 1, \ldots, n$ and are independent. Let L_1 be 1-dimensional subspace in $\mathbb{R}^n$, generated by unit-length vector a, $L_1 = \mathrm{Lin}(a)$. Let L_2 be a subspace orthogonal to L_1. Let T be a scaled ratio of lengths:

$$T = \frac{\langle a, z \rangle}{\|H_2 z\| / \sqrt{\dim L_2}}$$

where $\langle a, x \rangle$ is the length of the projection of vector z onto 1-dimensional subspace L_1 multiplied by plus or minus one, $\|H_2 z\|$ — the length of the projection of vector z onto L_2, Then T follows Student's distribution with $\dim L_2$ degrees of freedom.

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

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

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

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, $\langle a, z \rangle^2 = \|H_1 z\|^2 \sim \chi^2_1$ and $\|H_2 z\|^2 \sim \chi^2_k$. 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:

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

The latter equality can be illustarted with a 3-dimensional example (see Figure 36).

t-test

In a simple linear regression model

$$y = \beta_1 \mathbf{1} + \beta_2 x + \varepsilon$$

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

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

$$t = \frac{\hat{\beta} - \beta}{se\left(\hat{\beta}\right)}$$

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

$$t = \frac{\hat{\beta}_2}{se\left(\hat{\beta}_2\right)} \qquad (6)$$

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

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

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

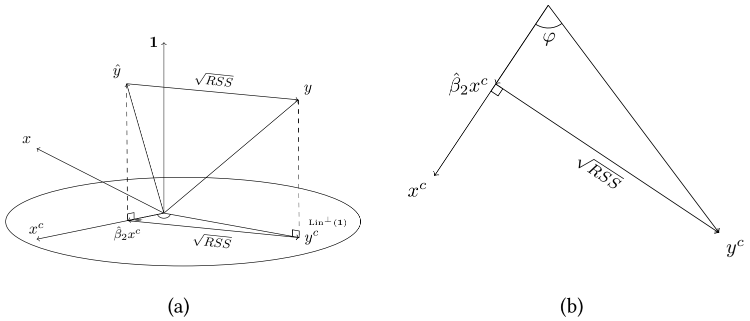
$$\begin{aligned}\frac{t}{\sqrt{n-2}} &= \frac{\hat{\beta}_2}{\sqrt{n-2}se\left(\hat{\beta}_2\right)} = \frac{\hat{\beta}_2}{\sqrt{n-2}\frac{\hat{\sigma}}{\sqrt{\sum\limits_{i=1}^{n}(x_i-\bar{x})^2}}} \\ &= \frac{\hat{\beta}_2\sqrt{\sum\limits_{i=1}^{n}(x_i-\bar{x})^2}}{\sqrt{n-2}\frac{\sqrt{\sum\limits_{i=1}^{n}(y_i-\hat{y}_i)^2}}{\sqrt{n-2}}} = \frac{\hat{\beta}_2\|x^c\|}{\sqrt{RSS}}\end{aligned}$$

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 $\operatorname{Lin}^{\perp}(\mathbf{1})$ so as to get their centred versions x^c and y^c. Then, we perform regression of y onto $\operatorname{Lin}(x, \mathbf{1})$ which results in $\hat{y}$. Following that, we project $\hat{y} = \hat{\beta}_1\mathbf{1} + \hat{\beta}_2 x$ onto $\operatorname{Lin}^{\perp}(\mathbf{1})$ which yields $\hat{\beta}_2 x^c$. After all, we translate $\sqrt{RSS}$ onto $\operatorname{Lin}^{\perp}(\mathbf{1})$. These steps are demonstrated in Figure 35(a).

Looking at Figure 35(b) which depicts the $\operatorname{Lin}^{\perp}(\mathbf{1})$, we derive

$$\operatorname{ctg}\varphi = \frac{\hat{\beta}_2\|x^c\|}{\sqrt{RSS}} = \frac{t}{\sqrt{n-2}}$$



(a) (b)

Figure 35: (a): Regression of y onto $\operatorname{Lin}(x, \mathbf{1})$ and appropriate projections; (b): $\operatorname{Lin}^{\perp}(\mathbf{1})$.

F-distribution

Definition 9. 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

$$F = \frac{\|H_1 z\|^2 / \dim L_1}{\|H_2 z\|^2 / \dim L_2} \sim F_{\dim L_1, \dim L_2},$$

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

Generally, the following definition is given.

Definition 8. Let $\gamma_1 \sim \chi^2_{k_1}$, $\gamma_2 \sim \chi^2_{k_2}$, γ_1, γ_2 independent. Then

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

Recall that by Theorem 14 the projections of a standard noraml vector onto orthogonal subspaces are independent. Thus, in terms of the Definition 9 $H_1 z$ and $H_2 z$ are independent. Next, from the Definition 6 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, $\|H_1 z\|^2 \sim \chi^2_{\dim L_1}$, $\|H_2 z\|^2 \sim \chi^2_{\dim L_2}$.

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:

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

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

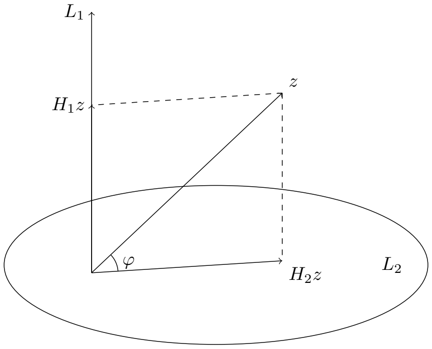


Figure 36: F-distribution as the ratio of the projection lengths squared adjusted to the dimensions of the subspaces.

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

$$y = \beta_1 \mathbf{1} + \beta_2 x + u$$

and the restricted model to be

$$y = \alpha_1 \mathbf{1} + v$$

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 37. Adjusted to the degrees of freedom, the ratio can be expressed in terms of the angle between two vectors, φ, as demonstrated in Figure 37

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

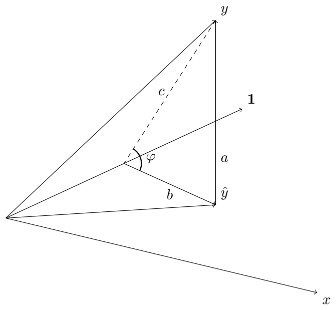


Figure 37: F-statistic is proportional to the cotangent squared of φ, where a stands for $\sqrt{RSS_{UR}}$, $b - \sqrt{RSS_R - RSS_{UR}}$, $c - \sqrt{RSS_R}$.

References

[1] George W Cobb. “Teaching statistics: Some important tensions”. In: *Chilean Journal of Statistics* 2.1 (2011), pp. 31–62.

[2] Erik Daniel Jacobson. “The Geometry of the General Linear Model”. University of Georgia, 2011.

[3] Leigh J. Halliwell. *The Gauss-Markov Theorem: Beyond the BLUE*. 2015. Paper presented at CAS E-Forum, fall 2015.

[4] Leandro da Silva Pereira, Lucas Monteiro Chaves, and Devanil Jaques de Souza. “An Intuitive Geometric Approach to the Gauss Markov Theorem”. In: *The American Statistician* 71.1 (2017), pp. 67–70. DOI: 10.1080/00031305.2016.1209127. eprint: https://doi.org/10.1080/00031305.2016.1209127. URL: https://doi.org/10.1080/00031305.2016.1209127.

[5] Richard J. Butler. “The Simple Geometry of Correlated Regressors and IV Corrections”. In: *International Journal of Statistics in Medical Research* 5.3 (2016), pp. 182–188. URL: http://dx.doi.org/10.6000/1929-6029.2016.05.03.4.

[6] Walter Sosa Escudero. *A Geometric Representation of the Frisch-Waugh-Lovell Theorem*. Department of Economics, Working Papers. 2001. URL: https://EconPapers.repec.org/RePEc:lap:wpaper:029.

[7] Øyvind Langsrud. “The geometrical interpretation of statistical tests in multivariate linear regression”. In: *Statistical Papers* 45.1 (Jan. 2004), pp. 111–122. ISSN: 1613-9798. DOI: 10.1007/BF02778273. URL: https://doi.org/10.1007/BF02778273.

[8] Enis Siniksaran. “On the geometry of F, Wald, LR, and LM tests in linear regression models”. In: *Statistics* 39.4 (2005), pp. 287–299. DOI: 10.1080/02331880500178521. eprint: https://doi.org/10.1080/02331880500178521. URL: https://doi.org/10.1080/02331880500178521.

[9] Michael Friendly, Georges Monette, and John Fox. “Elliptical Insights: Understanding Statistical Methods through Elliptical Geometry”. In: *Statist. Sci.* 28.1 (Feb. 2013), pp. 1–39. DOI: 10.1214/12-STS402. URL: https://doi.org/10.1214/12-STS402.