Gregory Kordas

Last update: May 2, 2023

LECTURE 5
LEAST SQUARES

1. A Historical Prelude

Isaac Newton's *Theory of Universal Gravitation*, introduced in his epoch-making book *Philosophia Naturalis Principia Mathematica*, often referred to as simply the *Principia*, had many wide-ranging implications about the physical world we live in. Greeted initially with incredulity and sometimes even downright mockery, the idea that masses mysteriously attract each other over vast distances of space seemed to many of his contemporaries counter intuitive and far fetched. Philosophical issues aside, the theory provided a wealth of predictions (about the motion and the shape of the planets, the tides of the seas, and the dark side of the Moon, to name just a few) that could be checked against empirical evidence and ultimately verify or falsify its premises.

One of the implications of Newton's theory was that, due to gravity, the rotation of the Earth around its axis would cause the Earth to bulge at the equator and flatten at the poles. More precisely, Newton proved that a rotating self-gravitating fluid body in equilibrium takes the form of an oblate ellipsoid of revolution (a spheroid). The exact amount of flattening depends on the body's density, its rotational speed, and the balance between the resulting gravitational and centrifugal forces. If gravity is therefore operational, it is unlikely that the Earth is a perfect sphere, but it should be an *oblate* spheroid, much like an orange.

Newton's theory was not the only theory around purporting to explain the motion of the celestial spheres. The French mathematician and physicist René Descartes had proposed the competing *Theory of Vortices.* According to Descartes' theory, space is filled with an invisible substance called the *ether*, that, much like water, creates vortices that sweep the planets into their apparent orbits. Now, plastic spherical objects inside a water vortex tend to flatten at the equator and bulge at the poles, so if the theory of Vortices was correct, the Earth should be a *prolate* spheroid, i.e., more like an egg or a lemon instead of an orange.

The two competing theories led to a prolonged controversy, much along nationalistic lines, between the English and their Continental rivals. Voltaire (1694 – 1778), who happened to be visiting London when Newton died in 1727, was greatly impressed by the State funeral and the honors bestowed on the great scientist and, on the subject of the controversy, commented that:

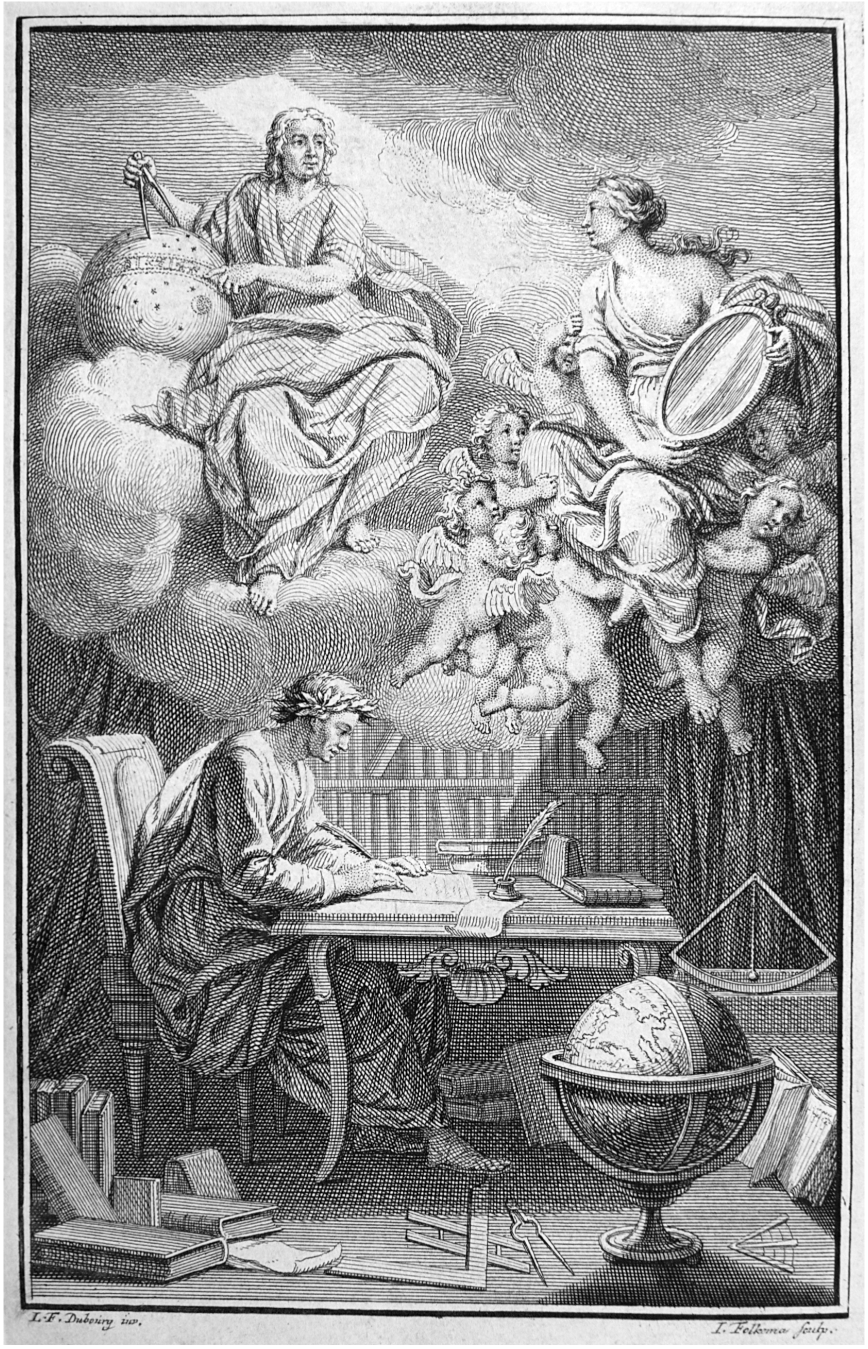

FIGURE 1. Frontispiece to Voltaire's book on Newton's philosophy. A Muse reflects Newton's heavenly insights down to Voltaire.

> A Frenchman arriving in London finds things very different. [...] For us it is the pressure of the Moon that causes the tides of the sea; for the English it is the sea that gravitates towards the Moon. [...] In Paris you see the Earth shaped like a melon, in London it is flattened out on two sides.

The equation of a 3D ellipsoid centred at the origin with semi-axes a, b and c aligned along the coordinate axes is

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = 1.$$

Because planets revolve around their north-south axis, physical considerations dictate that planets are *solids of revolution*, i.e., solids that can be obtained by revolving a 2D curve around the z (North-South) axis to obtain a 3D body. This means that for planets $a = b$, so the equation becomes

$$\frac{x^2 + y^2}{a^2} + \frac{z^2}{c^2} = 1.$$

If $c < a$ the ellipsoid is called *oblate* (orange-like), while if $c > a$ the ellipsoid is *prolate* (lemon-like). Of course, if $c = a$ we get a perfect sphere. Our interest is in the quantity of polar *flattening* or *ellipticity* f given by

$$f = \frac{a - c}{a} = 1 - \frac{c}{a}.$$

Letting $a = R_E$ be the Earth's equatorial radius and $c = R_P$ be its polar radius, we can write

$$f = 1 - \frac{R_P}{R_E}.$$

The Earth is oblate if $f > 0$, prolate if $f < 0$, and spherical if $f = 0$.

To settle the dispute once and for all, in 1735 the *Académie des Sciences Française* send expeditions to Ecuador, Lapland, and South Africa to measure meridians at widely separated latitudes. Along with the pre-existing measurements from Paris and Rome, the *Académie* managed to collect the five data points given in Table 1 (Stigler, 1986).

The length of 1° of latitude ℓ at the various locations was measured in toise, a popular measure of the time. To get a better idea, the table also presents these lengths in kilometers. A simple inspection of the table makes it apparent that the length of 1° of latitude grows as we move from the equator ($\theta = 0°$) to the poles ($\theta = 90°$). At Quito, which is on the equator, the length of a degree was measured to be 110.551 km, while at Lapland, which is the closest people of the time could get to the north pole on account of the cold, the length of the degree was measured to be more than a kilometer longer. Clearly these data favor Newton's prediction that the Earth flattens at the poles. There are, however, discrepancies too: at the Cape of

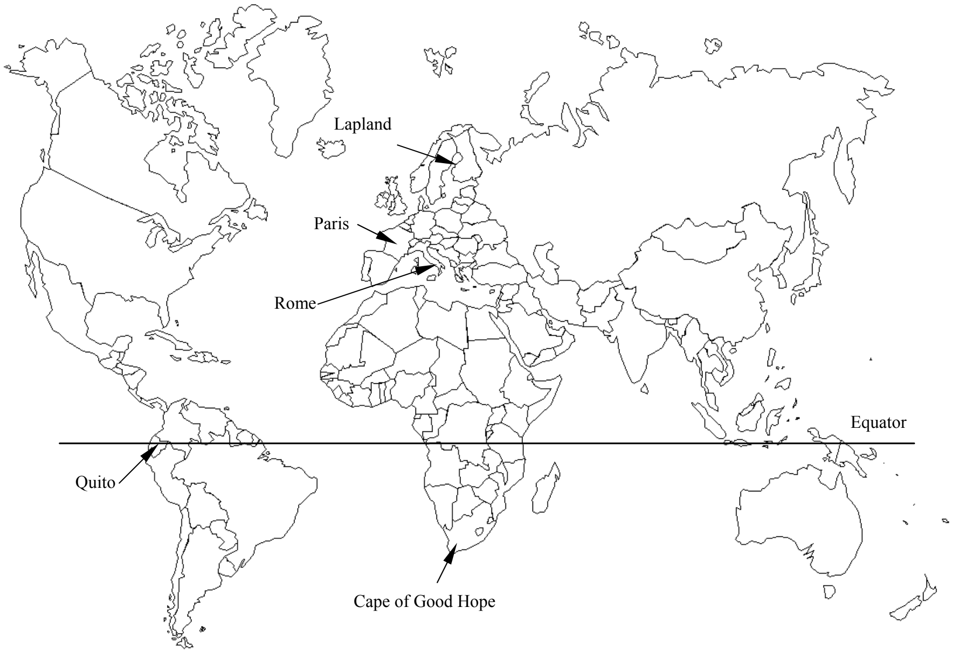

Figure 2. Location of available measurements on the world map.

Table 1. Data on the length of 1° of latitude at various locations.

Location	*Latitude* θ		$\sin^2(\theta)$	*Length of* 1° ℓ *(in toise)*	*Length of* 1° ℓ *(in km)*
Quito, Ecuador	0°	0′	0	56,751	110.551
Cape of Good Hope, S.Africa	33°	18′	0.2987	57,037	111.108
Rome, Italy	42°	59′	0.4648	56,979	110.995
Paris, France	49°	23′	0.5762	57,074	111.180
Lapland, Finland	66°	19′	0.8386	57,422	111.858

Note: 1 toise = 1.948 meters.

Good Hope the length of a degree is longer than that at Rome, despite the fact that Rome has a larger (north) latitude than the (south) latitude of the Cape of Good Hope. Graphing these measurements we see that, with the exception of the (Cape of Good Hope - Rome) pair, there seems to be a consistent tendency for the length to grow as we move away from the equator.

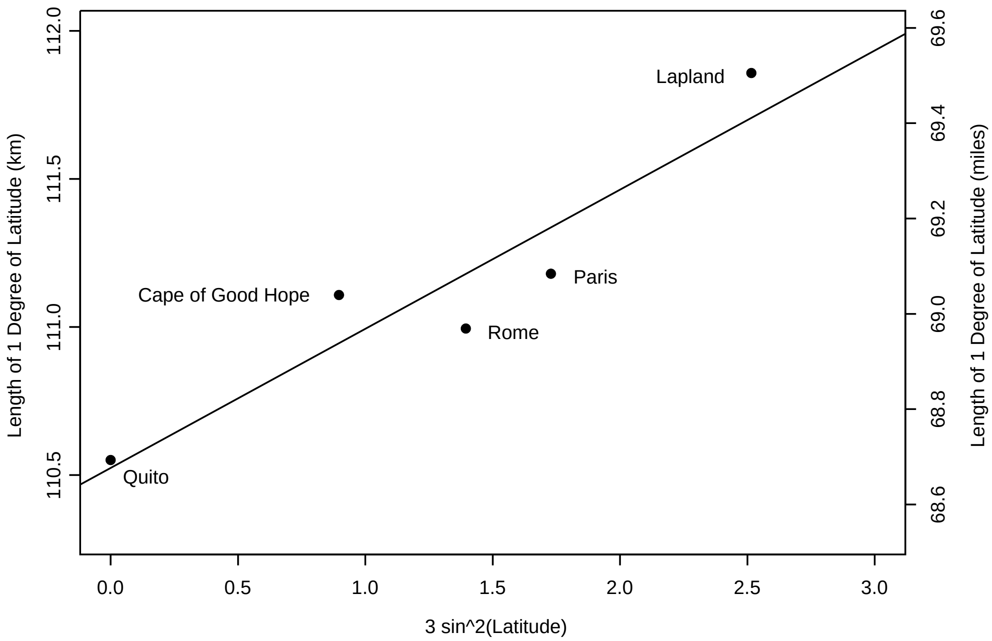

Figure 3. Graph of ℓ against $3\sin^2\theta$, along with the least squares line.

For short arcs, the approximation (see Stigler, 1986) (Nievergelt Yves (2001) A tutorial history of least squares with applications to astronomy and geodesy, in Brezinski C., L. Wuytack, Numerical Analysis (2001) – Historical Developments in the 20th Century (2001) p.77)

$$\ell = \beta_0 + \beta_1(3\sin^2\theta) + \text{higher-order terms},$$

where ℓ is the length of 1° of latitude and θ is the angle of the latitude, was known to be satisfactory. The parameters β_0 and β_1 can be interpreted as the length of 1° of latitude at the equator and the excess in length of 1° at the poles over its value at the equator, respectively. Ellipticity is therefore given by $f = \beta_1/\beta_0$.

These parameters are, of course, "known" today. Table 2 presents the geodetic constants for the *International Hayford Spheroid* (IHS). We see that the Earth flattens at the poles with an average oblate ellipticity of $f = 1/297$. The length of 1° of latitude at the equator is $60 \times 1,842.925 = 110,576$ m, while that at the poles is $60 \times 1,861.666 = 111,700$ m. This means that the circumference of the Earth around the equator is approximately[1] $C_E = 39,807$

[1] That the circumference of the earth in km's is almost a round number (40,000 km) is *not* a coincidence. The *meter* was originally defined as the one ten-millionth (1/10,000,000) of the distance between a pole and the equator along a great circle over water. Since to go around the earth one has to travel 4 times this distance, the circumference of the earth is 40 million meters.

TABLE 2. Geodetic Constants – Hayford International Spheroid$^{(*)}$.

$a = 6,378,388\,m$; $c = 6,356,912\,m$; $f = 1/297$

Latitude	*Length of 1′ of Longitude*	*Length of 1′ of Latitude*	*Acceleration of Gravity g*
	meters	*meters*	m/s^2
0°	1,855.398	1,842.925	9.780 350
15	1,792.580	1,844.170	9.783 800
30	1,608.174	1,847.580	9.793 238
45	1,314.175	1,852.256	9.806 154
60	930.047	1,856.951	9.819 099
75	481.725	1,860.401	9.828 593
90	0	1,861.666	9.832 072

$^{(*)}$Abramowitz M., and Stegun, I. A., (1972), *Handbook of Mathematical Tables*, New York: Dover, p.8.

km, while that around the poles is approximately $C_P = 39,807 \times (1 - 1/297) = 39,673$ km, a mere 134 km less than that around the equator.

In Book III of the *Principia*, Newton himself predicted that

> [...] the diameter of the Earth at the equator is to its diameter from pole to pole as 230 to 229. – *Principia*, Book III, Proposition XIX, Problem III.

That is, Newton gave $f = 1/230$. Fitzpatrick (2009, sec. 2.12)[2] presents a theoretical model of rotational flattening that, under simplifying homogeneity assumptions about the rotating body, predicts $f = 1/233$. He says that this is (essentially) the model that Newton used to make his prediction, and comments that "the discrepancy [with the actual $f = 1/297$ value] is due to the fact that the Earth is strongly inhomogeneous, being much denser at its core than in its outer regions".

In terms of our model, $\beta_0 = 110.576$ km, while from f we obtain a polar exceedance of $\beta_1 = 110.576/297 = 0.3723$ km per 1° of latitude. In what follows, we will estimate β_0, β_1, and f from the data in Table 1 and see how close the scientists of the time came to discovering the "truth".

[2] Fitzpatrick R. (2009) *Theoretical Fluid Mechanics*, University of Texas at Austin Press.

2. Linear Models

Based on theoretical considerations, a scientist believes that there is a *linear function* relating a scalar quantity y to a k vector of variables $\boldsymbol{x}$. The function itself is *known only up to a finite set of k parameters* $\boldsymbol{\beta}$ that he wishes to estimate from data. For this purpose, a *sample* is obtained, given by

$$\{(y_1, \boldsymbol{x}_1), (y_2, \boldsymbol{x}_2), ..., (y_n, \boldsymbol{x}_n)\} = \{(y_i, \boldsymbol{x}_i) : i = 1, ..., n\}$$

where each pair $(y_i, \boldsymbol{x}_i) \in \mathbb{R} \times \mathbb{R}^k$. In terms of the i-th observation, the linear model is given by

$$y_i = \boldsymbol{x}_i' \boldsymbol{\beta} + u_i, \qquad i = 1, ..., n,$$

where y_i is the i-th value of the *dependent* variable, $\boldsymbol{x}_i = (x_{1i}, x_{2i}, ..., x_{ki})'$ with $x_{1i} \equiv 1$ is the $k \times 1$ vector containing the values of the *independent* variables for the i-th observation in the sample, and u_i is a *random disturbance*. For notational purposes, it is often convenient to arrange the observations in matrix form as follows:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1k} \\ x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix},$$

or

$$\boldsymbol{y}_{n \times 1} = X_{n \times k} \boldsymbol{\beta}_{k \times 1} + \boldsymbol{u}_{n \times 1},$$

where $\boldsymbol{y}$ is an $n \times 1$ vector, $\boldsymbol{X}$ is a $n \times k$ matrix, $\boldsymbol{\beta}$ is a $k \times 1$ vector of coefficients to be estimated, and $\boldsymbol{u}$ is a $n \times 1$ vector of residuals.

In terms of our discussion above, $y_i = \ell_i$ which, for short arcs, can be sufficiently approximated by a linear function of $\boldsymbol{x}_i = (1, 3\sin^2\theta_i)'$. The residual term u_i represents the scientist's ignorance regarding several "unobserved" components:

(i) First, it incorporates "higher-order" terms of the *true functional equation* connecting y_i to $\boldsymbol{x}_i$. In terms of our example, the linear equation connecting ℓ to θ is only valid to a first approximation. The true relation between ℓ and θ is indeed quite complicated, but we hope that, at least for short-arcs, the linear equation is sufficiently accurate.

(ii) Second, the residual term incorporates deviations between the idealizations of a our mathematical model and reality. In terms of our example, the mathematical model employed assumes that the Earth is a perfect smooth spheroid that can be described by a *single* ellipticity parameter f! In reality, of course, the Earth is not a perfect spheroid, but has a very complicated surface with bulges and depressions (mount Everest is 8.85 km above sea level, and the Mariana Trench is 10.91 km below sea level), as well as, asymmetries (the Earth has more mass on the Northern than on the Southern Hemisphere, so, to a second degree approximation, it looks like a potato).

(iii) Finally, the residual incorporates observational noise. The observations are noisy due to the limited precision of our instruments and various other random factors that affect the measurements.

In what follows we will discuss the Ordinary Least Squares (OLS) estimates of $\boldsymbol{\beta}$, emphasizing three alternative ways of interpreting these estimates.

3. Least Squares as a Solution to an Over-Identified System of Equations

The hypothesized linear model evaluated at the n data points produces a system of n equations with k unknowns. More specifically, the system can be written as

$$\left.\begin{array}{lcl} y_1 & = & x_{11}\beta_1 + x_{21}\beta_2 + \cdots + x_{k1}\beta_k \\ y_2 & = & x_{12}\beta_1 + x_{22}\beta_2 + \cdots + x_{k2}\beta_k \\ \cdots & \cdots & \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ y_n & = & x_{1n}\beta_1 + x_{2n}\beta_2 + \cdots + x_{kn}\beta_k \end{array}\right\} \tag{3.1}$$

with $x_{1i} \equiv 1$.

If the equations are linearly independent and $n > k$ (as it is always the case if we are diligent enough to collect enough data), then the system is over-identified, and has no solution, i.e., there is no signle k-vector $\boldsymbol{\beta}$ that can simultaneously satisfy all n equations. We could throw away the excess equations (data points) so as to make our system exactly identified, but this would certainly not be an optimal strategy: if the data are noisy then combining all the available information would definitely be preferable as it would produce less noisy estimates.[3] Another idea would be to note that a system like this contains $\binom{n}{k}$ k-equation systems that each has a unique solution, so perhaps it would be a good idea to compute them all. Figure 3 presents all the $\binom{5}{2} = 10$ possible lines that go through each pair of points in our application. However, this doesn't look like a very good idea either, since it produces too many "solutions" without reducing the noise in the data.

[3] This argument, that combining errors tends to reduce the overall error of the estimate, was very difficult to gain acceptance in the early days of Statistics, since mathematicians generally believe that errors pile up and do not cancel out. Of course, the point is that *systematic* errors pile up, but *random* errors (noise) tend to cancel each other out.

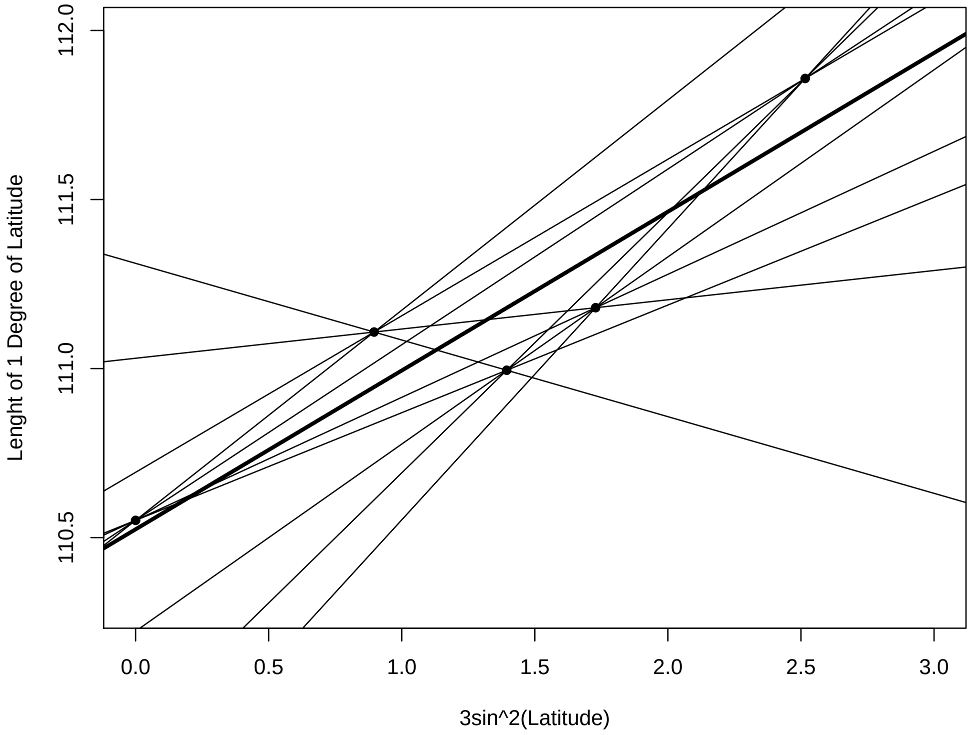

Figure 4. The $\binom{5}{2} = 10$ possible lines (thin) along with the least squares line (bold).

The key to solving our problem is to note that the noise in the data would be reduced if we were to somehow *combine* the $\binom{n}{k}$ solutions into a single one. This can be done by specifying a *loss function*, i.e., a function that quantifies our intention to fit a single equation through the data in a way that the residuals are "minimal". We have already seen square and absolute loss functions. The reason that people favoured the square loss initially is simply because it is easier to work with: both the square and absolute loss functions are is globally convex, but the square is also differentiable, a convenient property that the absolute loss doesn't share (the absolute loss $|u|$ does not have a derivative at $u = 0$).

The problem, therefore, is to minimize the squared deviations from the observed y, i.e. solve

$$\begin{aligned}
\min_{\boldsymbol{\beta} \in \mathbb{R}^k} S(\boldsymbol{\beta}) &= \boldsymbol{u}'\boldsymbol{u} \\
&= (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \\
&= \boldsymbol{y}'\boldsymbol{y} - \boldsymbol{y}'\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{y} + \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} \\
&= \boldsymbol{y}'\boldsymbol{y} - 2\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{y} + \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta},
\end{aligned}$$

where the last equality follows from that $\boldsymbol{y}'\boldsymbol{X}\boldsymbol{\beta}$ is a scalar and thus equal to its transpose $\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{y}$. The following Aside reminds us how to differentiate linear and quadratic forms.

Aside. In differentiating linear $(\boldsymbol{a}'\boldsymbol{x}, \boldsymbol{A}\boldsymbol{x})$ and quadratic $(\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}, \boldsymbol{y}'\boldsymbol{A}\boldsymbol{x})$ forms, the following rules hold (we assume that all vectors and matrices below are conformable and the products are defined):

(i) $\dfrac{\partial \boldsymbol{a}'\boldsymbol{x}}{\partial \boldsymbol{x}} = \boldsymbol{a}'$ and $\dfrac{\partial \boldsymbol{x}'\boldsymbol{a}}{\partial \boldsymbol{x}} = \boldsymbol{a}$

(ii) $\dfrac{\partial \boldsymbol{A}\boldsymbol{x}}{\partial \boldsymbol{x}} = \boldsymbol{A}'$ and $\dfrac{\partial \boldsymbol{x}'\boldsymbol{A}}{\partial \boldsymbol{x}} = A$

(iii) $\dfrac{\partial \boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}}{\partial \boldsymbol{x}} = 2\boldsymbol{x}'\boldsymbol{A}$ [$= 2\boldsymbol{A}\boldsymbol{x}$ if $\boldsymbol{A}$ is symmetric]

(iv) $\dfrac{\partial \boldsymbol{y}'\boldsymbol{A}\boldsymbol{x}}{\partial \boldsymbol{x}} = \boldsymbol{y}'\boldsymbol{A}$, [assuming $\boldsymbol{A}$ is symmetric]

(v) $\dfrac{\partial \boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}}{\partial \boldsymbol{A}} = \boldsymbol{x}\boldsymbol{x}'$

(vi) $\dfrac{\partial \boldsymbol{y}'\boldsymbol{A}\boldsymbol{x}}{\partial \boldsymbol{A}} = \boldsymbol{x}\boldsymbol{y}'$, [assuming $\boldsymbol{A}$ is symmetric]

■

Differentiating and setting the derivative equal to zero, we obtain the so called *normal equations*

$$\frac{\partial S(\hat{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}} = -2\boldsymbol{X}'\boldsymbol{y} + 2\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}} = 0.$$

Provided that the $k \times k$ matrix $\boldsymbol{X}'\boldsymbol{X}$ is of full rank and can thus be inverted, we solve for $\hat{\boldsymbol{\beta}}$ to obtain the *OLS coefficients*,

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} \\ &= \left(\sum_{i=1}^{n} \boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1} \sum_{i=1}^{n} \boldsymbol{x}_i y_i \,. \end{aligned} \tag{3.2}$$

This is the solution that the minimization of the square loss function $S(\boldsymbol{\beta}) = \boldsymbol{u}'\boldsymbol{u}$ produces for the over-identified system (3.1). Note that $\hat{\mathcal{M}}_{\boldsymbol{xx}} \equiv \boldsymbol{X}'\boldsymbol{X}$ is the sample covariance matrix of the regressors $\boldsymbol{x}$, and $\hat{\mathcal{M}}_{\boldsymbol{x}y} \equiv \boldsymbol{X}'\boldsymbol{y}$ is the sample covariance of the regressors $\boldsymbol{x}$ with the dependent variable y, so

$$\hat{\boldsymbol{\beta}} = \hat{\mathcal{M}}_{\boldsymbol{xx}}^{-1}\hat{\mathcal{M}}_{\boldsymbol{xy}}. \tag{3.3}$$

In the simple regression model ($k = 2$, i.e., a model with a single regressor and an intercept), we have

$$(\boldsymbol{X}'\boldsymbol{X})^{-1} = \frac{1}{\sum x_i^2 - \left(\sum x_i\right)^2} \begin{bmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n \end{bmatrix}.$$

and the OLS coefficients are

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} \\ &= \frac{1}{n\sum x_i^2 - \left(\sum x_i\right)^2} \begin{bmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n \end{bmatrix} \begin{bmatrix} \sum y_i \\ \sum x_i y_i \end{bmatrix} \\ &= \frac{1}{n\sum x_i^2 - \left(\sum x_i\right)^2} \begin{bmatrix} \sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i \\ n\sum x_i y_i - \sum x_i \sum y_i \end{bmatrix}. \end{aligned}$$

The slope coefficient $\hat{\beta}_1$ is

$$\hat{\beta}_1 = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - \left(\sum x_i\right)^2} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2}$$

that can be written as

$$\hat{\beta}_1 = \frac{\sum (y_i - \bar{y})(x_i - \bar{x})}{\sum (x_i - \bar{x})^2} = \frac{\widehat{Cov}(y_i, x_i)}{\widehat{Var}(x_i)}, \tag{3.4}$$

where $\widehat{Cov}(y_i, x_i)$ is the sample covariance of y_i and x_i, and $\widehat{Var}(x_i)$ is that sample variance of x_i. Equation (3.4) is a special case of the general moment equation (3.3).

The constant $\hat{\beta}_0$ is

$$\begin{aligned} \hat{\beta}_0 &= \frac{(n\bar{y})\sum x_i^2 - (n\bar{x})\sum x_i y_i}{n\sum x_i^2 - (n\bar{x})^2} \\ &= \frac{\bar{y}\left[\sum x_i^2 - (n\bar{x})\right] + \bar{y}(n\bar{x})^2 - (n\bar{x})\sum x_i y_i}{n\sum x_i^2 - (n\bar{x})^2} \\ &= \bar{y} - \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2}\,\bar{x} \\ &= \bar{y} - \hat{\beta}_1 \bar{x}. \end{aligned} \tag{3.5}$$

where $\bar{y}$ and $\bar{x}$ are the sample means of the regressand y and the regressor x, respectively.

Aside. To see exactly how the least squares estimator $\hat{\boldsymbol{\beta}}$ combines the $\binom{n}{k}$ k-equation possible solutions, consider the simple regression model and let h index the $\binom{n}{k}$ pairs, and write

$$\boldsymbol{X}(h) = \begin{bmatrix} 1 & x_i \\ 1 & x_j \end{bmatrix}, \quad \boldsymbol{y}(h) = \begin{bmatrix} y_i \\ y_j \end{bmatrix}$$

where for the simple bivariate model $h = (i, j)$. Then since $\boldsymbol{X}(h)$ is a 2×2 square matrix and assuming that it is nonsingular

$$\boldsymbol{b}(h) = \boldsymbol{X}(h)^{-1}\boldsymbol{y}(h), \tag{3.6}$$

is the intercept and slope of the line passing through pair $h = (i, j)$. The following theorem shows that we may now write the least squares estimator as

$$\hat{\boldsymbol{\beta}} = \sum_h w(h)\boldsymbol{b}(h), \tag{3.7}$$

where

$$w(h) = \frac{|\boldsymbol{X}(h)|^2}{\sum_h |\boldsymbol{X}(h)|^2},$$

and $|\boldsymbol{X}(h)|$ is the determinant of $\boldsymbol{X}(h)$, and h ranges over the $\binom{n}{2}$ tuples. We see that the OLS estimate is a weighted average of the $\binom{n}{k}$ lines, with weights proportional to $|\boldsymbol{X}(h)|^2$, the square of the determinant of $\boldsymbol{X}(h)$. The result generalizes directly to the linear regression model with k regressors.

THEOREM 1. *(Subrahmanyam (1972))*[4] *The OLS estimator $\hat{\boldsymbol{\beta}}$ for the general linear regression model with k regressors can be written as in (3.6) and (3.7) for k-tuples $h = (i_1, i_2, ..., i_k)$. In the general case, $\boldsymbol{X}(h)$ is a $k \times k$ square matrix, $\boldsymbol{y}(h)$ is a $k \times 1$ vector, and h ranges over all the possible $\binom{n}{k} = O(n^k)$ k-tuples in the sample.*

Proof. Since $\boldsymbol{X}(h)$ is $k \times k$ square matrix,

$$|\boldsymbol{X}(h)'\boldsymbol{X}(h)| = |\boldsymbol{X}'(h)||\boldsymbol{X}(h)| = |\boldsymbol{X}(h)|^2.$$

The key to the proof of Subrahmanyam's result is a theorem from the theory of determinants[5] which states that

$$|\boldsymbol{X}'\boldsymbol{X}| = \sum_h |\boldsymbol{X}(h)|^2,$$

where h ranges over the $\binom{n}{k}$ combinations of n things taken k at a time.

[4] Hoerl, A. E. and Kennard, R. W. (1980), "M30. A note on least squares estimates", Communications in Statistics - Simulation and Computation, 9(3).

Subrahmanyam, M. (1972), "A Property of Simple Least Squares Estimates," Sankhya, Series B, Indian Journal of Statistics, 34, 355–356.

Wu, C. F. J. (1986), "Jackknife, Bootstrap and Other Resampling Methods in Regression Analysis," Annals of Statistics, 14, 1261–1295.

[5] See page 33, problem 2.6 of C.R. Rao (1973), *Linear Statistical Inference and Its Applications*, John Wiley and Sons.

Let $\boldsymbol{S} = \boldsymbol{X}'\boldsymbol{X}$. Denote by $\boldsymbol{X}_j$ the matrix obtained by replacing the j-th column of $\boldsymbol{X}$ by $\boldsymbol{y}$, so that $\boldsymbol{S}_j = \boldsymbol{X}'\boldsymbol{X}_j$. By Cramer's rule, the jth LS estimate for the hth k-tuple is given by

$$b_j(h) = \frac{|\boldsymbol{S}_j(h)|}{|\boldsymbol{S}(h)|} = \frac{|\boldsymbol{X}(h)'\boldsymbol{X}_j(h)|}{|\boldsymbol{X}(h)'\boldsymbol{X}(h)|} = \frac{|\boldsymbol{X}(h)|\,|\boldsymbol{X}_j(h)|}{|\boldsymbol{X}(h)|^2} = \frac{|\boldsymbol{X}_j(h)|}{|\boldsymbol{X}(h)|},$$

since $\boldsymbol{X}(h)$ and $\boldsymbol{X}_j(h)$ are square matrices, so that

$$|\boldsymbol{X}_j(h)| = b_j(h)\,|\boldsymbol{X}(h)|.$$

Again by Cramer's rule, the jth LS estimate $\hat{\beta}_j$ based on all the observations is given by

$$\begin{aligned}
\hat{\beta}_j &= \frac{|\boldsymbol{S}_j|}{|\boldsymbol{S}|} = \sum_h \frac{|\boldsymbol{S}_j(h)|}{|\boldsymbol{S}|} \\
&= \sum_h \frac{|\boldsymbol{X}(h)'\boldsymbol{X}_j(h)|}{|\boldsymbol{S}|} \\
&= \sum_h \frac{|\boldsymbol{X}(h)||\boldsymbol{X}_j(h)|}{|\boldsymbol{S}|} \qquad \text{since } \boldsymbol{X}(h) \text{ and } \boldsymbol{X}_j(h) \text{ are square matrices} \\
&= \sum_h \frac{|\boldsymbol{X}(h)|^2}{\sum_h |\boldsymbol{X}(h)|^2}\, b_j(h) \\
&= \sum_h w(h)\, b_j(h).
\end{aligned}$$

□

An interesting implication of this result is that the $\boldsymbol{b}(h)$'s that correspond to k-tuples of observations that produce singular $\boldsymbol{X}(h)$ matrices, receive zero weight in the LS estimator $\hat{\boldsymbol{\beta}}$ using all the n observations. ■

The *OLS fitted values* are given by

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} \equiv \boldsymbol{P_X}\boldsymbol{y},$$

and the *OLS residuals* by

$$\hat{\boldsymbol{u}} = \boldsymbol{y} - \hat{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{P_X}\boldsymbol{y} = (\boldsymbol{I} - \boldsymbol{P_X})\boldsymbol{y} \equiv \boldsymbol{M_X}\boldsymbol{y}.$$

It is clear that we can now decompose $\boldsymbol{y}$ into

$$\boldsymbol{y} = \hat{\boldsymbol{y}} + \hat{\boldsymbol{u}} = \boldsymbol{P_X}\boldsymbol{y} + \boldsymbol{M_X}\boldsymbol{y}.$$

The matrices $\boldsymbol{P_X}$ and $\boldsymbol{M_X}$ are called *projection matrices* and they have special properties. One way to think about them is to see them as *filters* that extract different parts out of $\boldsymbol{y}$:

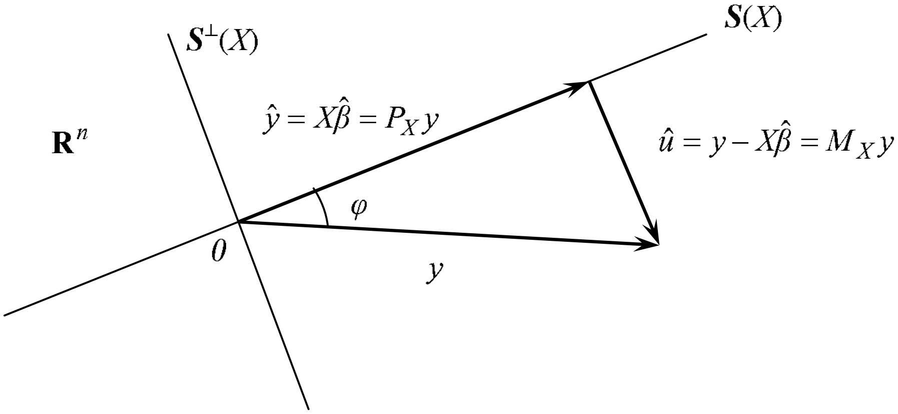
$S(X)$
$S^{\perp}(X)$
$\mathbf{R}^n$
$\hat{y} = X\hat{\beta} = P_X y$
$\hat{u} = y - X\hat{\beta} = M_X y$
φ
0
y

FIGURE 5. The Geometry of Least Squares.

when $\boldsymbol{P_X}$ is applied on $\boldsymbol{y}$ it extracts fitted values $\hat{\boldsymbol{y}}$ from it, while, when $\boldsymbol{M_X}$ is applied on $\boldsymbol{y}$ it extracts residuals $\hat{\boldsymbol{u}}$. But to understand these projection matrices better we need to shift our way of thinking and look at Least Squares from a *geometric* point of view.

4. LEAST SQUARES AS AN ORTHOGONAL PROJECTION

The vector $\boldsymbol{y}$ belongs to $\mathbb{R}^n$, and if $\boldsymbol{X}$ has rank $k < n$, the columns of $\boldsymbol{X}$ span a k-dimensional subspace of $\mathbb{R}^n$. Let $\mathcal{S}(\boldsymbol{X})$ be the (linear) subspace spanned by $\boldsymbol{X}$

$$\mathcal{S}(\boldsymbol{X}) = \left\{\boldsymbol{X\beta} : \boldsymbol{\beta} \in \mathbb{R}^k\right\},$$

and let $\mathcal{S}^{\perp}(\boldsymbol{X})$ be its orthogonal linear subspace. By construction, $\mathcal{S}(\boldsymbol{X})$ has dimension k and $\mathcal{S}^{\perp}(\boldsymbol{X})$ has dimension $n-k$, so, taken together, the two subspaces span the entire $\mathbb{R}^n$. This means that $\boldsymbol{y}$, which is a vector in $\mathbb{R}^n$, can be written as a linear combination of elements in $\mathcal{S}(\boldsymbol{X})$ and $\mathcal{S}^{\perp}(\boldsymbol{X})$. According to the standard terminology of linear algebra, $\mathcal{S}(\boldsymbol{X})$ is the *column space* of $\boldsymbol{X}$ and $\mathcal{S}^{\perp}(\boldsymbol{X})$ is the *null space* of $\boldsymbol{X}$, but for obvious reasons we will refer to them as the *subspace of fitted values* and the *subspace of residuals*, respectively.

Our objective is to find the element $\boldsymbol{X}\hat{\boldsymbol{\beta}}$ of $\mathcal{S}(\boldsymbol{X})$ that minimizes the distance between $\boldsymbol{y}$ and $\mathcal{S}(\boldsymbol{X})$. Let $\boldsymbol{u} = \boldsymbol{y} - \boldsymbol{X\beta}$ be the residual vector resulting from approximating $\boldsymbol{y}$ by $\boldsymbol{X\beta}, \boldsymbol{\beta} \in \mathbb{R}^k$. Our objective can thus be stated as finding the $\boldsymbol{u}$ with the *minimum length.* Geometrically, the residual vector with a minimum length $\hat{\boldsymbol{u}}$ is the vector that is *perpendicular* to $\mathcal{S}(\boldsymbol{X})$, i.e.

$$\hat{\boldsymbol{u}} \perp \mathcal{S}(\boldsymbol{X}).$$

It follows that $\hat{\boldsymbol{u}} \perp \boldsymbol{X}$ also, which implies that $\hat{\boldsymbol{u}}$ and $\boldsymbol{X}$ are uncorrelated, that is, $\boldsymbol{X}'\hat{\boldsymbol{u}} = \boldsymbol{0}$, from which it follows that

$$\begin{aligned} & \boldsymbol{X}'\hat{\boldsymbol{u}} = \boldsymbol{0} \\ \Leftrightarrow \quad & \boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{X}'\hat{\boldsymbol{\beta}}) = \boldsymbol{0} \\ \Leftrightarrow \quad & \boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}'\boldsymbol{y} \\ \Leftrightarrow \quad & \hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}. \end{aligned}$$

We have thus arrived at the OLS solution through a *geometrical argument.* It is perhaps worth noting that we would have arrived at the same solution had we chosen to set $\hat{\boldsymbol{u}}$ orthogonal to any other element of $\mathcal{S}(\boldsymbol{X})$: $\hat{\boldsymbol{u}} \perp \boldsymbol{X}\boldsymbol{c}, \boldsymbol{c} \in \mathbb{R}^k$, would have given us, $\boldsymbol{c}'\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) = 0 \Leftrightarrow \boldsymbol{c}'\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{c}'\boldsymbol{X}'\boldsymbol{y} \Leftrightarrow \hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$.

That the OLS solution has this geometrical interpretation should not be very surprising. Recalling that the *Euclidean length* of a vector $\boldsymbol{z}$ in $\mathbb{R}^n$ is given by

$$||\boldsymbol{z}|| = \sqrt{\sum_{i=1}^{n} z_i^2} = \sqrt{\boldsymbol{z}'\boldsymbol{z}}.$$

we see immediately that the minimizer $\hat{\boldsymbol{\beta}}$ of the length of the residual vector $\boldsymbol{u}$ is given by

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \; ||\boldsymbol{u}|| = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \; \sqrt{\boldsymbol{u}'\boldsymbol{u}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \; \boldsymbol{u}'\boldsymbol{u}$$

which is the same problem as the one we considered in the previous subsection.

The geometric properties of the OLS solution justify us calling $\boldsymbol{P_X}$ and $\boldsymbol{M_X}$ *projection matrices*: $\boldsymbol{P_X}$ orthogonally projects $\boldsymbol{y}$ into the “space of fitted values” $\mathcal{S}(\boldsymbol{X})$, and $\boldsymbol{M_X}$ orthogonally projects $\boldsymbol{y}$ into the “space of residuals” $\mathcal{S}^{\perp}(\boldsymbol{X})$.

A square matrix $\boldsymbol{A}$ is idempotent if “squaring it” leaves it unchanged, i.e., $\boldsymbol{A}\boldsymbol{A} = \boldsymbol{A}$. Observing that $\boldsymbol{P_X} = \boldsymbol{P}'_{\boldsymbol{X}}$, $\boldsymbol{M_X} = \boldsymbol{M}'_{\boldsymbol{X}}$ and

$$\begin{aligned} \boldsymbol{P_X}\boldsymbol{P_X} &= [\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'][\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'] = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' = \boldsymbol{P_X} \\ \boldsymbol{M_X}\boldsymbol{M_X} &= [\boldsymbol{I} - \boldsymbol{P_X}][\boldsymbol{I} - \boldsymbol{P_X}] = \boldsymbol{I} - 2\boldsymbol{P_X} + \boldsymbol{P_X}\boldsymbol{P_X} = \boldsymbol{I} - \boldsymbol{P_X} = \boldsymbol{M_X}, \end{aligned}$$

we see that $\boldsymbol{P_X}$ and $\boldsymbol{M_X}$ are symmetric idempotent matrices (in fact all projection matrices are idempotent, although not necessarily symmetric – see the section on restricted least squares). The reason for this peculiar behavior of $\boldsymbol{P_X}$ and $\boldsymbol{M_X}$ can best be understood geometrically. When we apply $\boldsymbol{P_X}$ to a vector $\boldsymbol{y}$ we project it onto $\mathcal{S}(\boldsymbol{X})$. Now re-applying $\boldsymbol{P_X}$ on $\boldsymbol{P_X}\boldsymbol{y}$ leaves the vector $\boldsymbol{P_X}\boldsymbol{y}$ unchanged since $\boldsymbol{P_X}\boldsymbol{y}$ already belongs to $\mathcal{S}(\boldsymbol{X})$ and the second projection does nothing. Likewise, applying $\boldsymbol{M_X}$ on $\boldsymbol{y}$ carries $\boldsymbol{y}$ into $\mathcal{S}^{\perp}(\boldsymbol{X})$ but re-applying $\boldsymbol{M_X}$ on the already transformed vector $\boldsymbol{M_X y}$ leaves it unchanged.

Another property of the matrices $\boldsymbol{P_X}$ and $\boldsymbol{M_X}$ is that they are orthogonal to each other, a fact that is also obvious geometrically. In fact, for any $n \times 1$ vector $\boldsymbol{z}$, $\boldsymbol{M_X z}$ and $\boldsymbol{P_X z}$ are orthogonal, exactly because these matrices project $\boldsymbol{z}$ onto subspaces that are orthogonal to each other. Algebraically, we have the following trivial result

$$\boldsymbol{P_X M_X} = \boldsymbol{P_X}[\boldsymbol{I} - \boldsymbol{P_X}] = \boldsymbol{P_X} - \boldsymbol{P_X P_X} = \boldsymbol{0},$$

where $\boldsymbol{0}$ is a $k \times k$ matrix of zeros.

Aside. Below we list some very useful theorems regarding idempotent matrices that will be needed in the succeeding sections [6].

THEOREM A. If A is an $n \times n$ symmetric matrix of rank k, then a necessary and sufficient condition that A is idempotent is that each of k of the characteristic roots of A is equal to unity and the remaining $(n-k)$ characteristic roots are equal to zero, i.e., $\lambda_1 = \cdots = \lambda_k = 1$ and $\lambda_{k+1} = \cdots = \lambda_n = 0$.

THEOREM B. If A is an idempotent matrix, then rank $A =$ trace A.

THEOREM C. The only nonsingular idempotent matrix is the identity matrix.

THEOREM D. If A is an $n \times n$ idempotent matrix of rank k such that $k < n$ $(k = n)$, then A is a positive semidefinite matrix (positive definite matrix).

THEOREM E. If A is an idempotent matrix whose ith diagonal element is equal to zero, then every element in the ith row and ith column of A is equal to zero.

■

[6] These theorems are from Graybill, F.A, and Marsaglia G. (1957) – Idempotent Matrices and Quadratic Forms in the General Linear Hypothesis, *Annals of Mathematical Statistics*, vol. 28, pp. 678-686)

Recalling that $\boldsymbol{y} = \hat{\boldsymbol{y}} + \hat{\boldsymbol{u}}$, we can decompose the sum of squares of $\boldsymbol{y}$, $||\boldsymbol{y}||^2$, as

$$\begin{aligned} \boldsymbol{y}'\boldsymbol{y} &= (\hat{\boldsymbol{y}} + \hat{\boldsymbol{u}})'(\hat{\boldsymbol{y}} + \hat{\boldsymbol{u}}) \\ &= (\boldsymbol{P_X y} + \boldsymbol{M_X y})'(\boldsymbol{P_X y} + \boldsymbol{M_X y}) \\ &= \boldsymbol{y}'\boldsymbol{P_X y} + \boldsymbol{y}'\boldsymbol{M_X y} \\ &= ||\boldsymbol{P_X y}||^2 + ||\boldsymbol{M_X y}||^2. \end{aligned}$$

where $||\boldsymbol{z}||$ again denotes the Euclidean length of the vector $\boldsymbol{z}$. This is nothing more than Pythagoras' theorem: it terms of Figure 5, the square of the hypotenuse equals the sum of squares of the sides of the triangle traced by $\boldsymbol{y}$, $\boldsymbol{P_X y}$ and $\boldsymbol{M_X y}$. This decomposition gives us the *uncentered-*R^2

$$R_u^2 = \frac{||\boldsymbol{P_X y}||^2}{||\boldsymbol{y}||^2} = 1 - \frac{||\boldsymbol{M_X y}||^2}{||\boldsymbol{y}||^2}.$$

It is clear that R_u^2 is a unit-free number that takes values between 0 and 1. It is also simple to see that it too has a simple geometric interpretation: it is the square of the cosine of the angle between the vectors $\boldsymbol{y}$ and $\boldsymbol{P_X y}$, i.e.,

$$R_u^2 = \cos^2 \varphi.$$

Unfortunately, however, R_u^2 is not entirely satisfactory in that it measures as "fit", both the effect of the intercept as well as the rest of the x's in $\boldsymbol{X}$. In applications, we are interested in a measure of fit that tells us how much of the total variation of $\boldsymbol{y}$ is explained by the variable regressors, not the intercept. Recall that if we regress $\boldsymbol{y}$ on an intercept alone, that is, if we regress $\boldsymbol{y}$ on $\mathbf{1}$, where $\mathbf{1} = (1, 1, ..., 1)'$ is a $n \times 1$ vector of ones, we will obtain the mean of $\boldsymbol{y}$,

$$(\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}'\boldsymbol{y} = \mathbf{1}'\boldsymbol{y}/n = \bar{y}.$$

The *centered-*R^2 is defined as the sum of squares of $\boldsymbol{y}$ explained by $\boldsymbol{X}$ once we remove the effect of the intercept by de-meaning $\boldsymbol{y}$. To de-mean $\boldsymbol{y}$, we project it onto the space spanned by $\mathbf{1}$, and obtain the residual $\boldsymbol{M_1 y}$. The centered R^2 used in applications is then given by

$$R^2 = 1 - \frac{||\boldsymbol{M_X y}||^2}{||\boldsymbol{M_1 y}||^2}$$

where

$$\boldsymbol{M_1} = \boldsymbol{I} - \mathbf{1}(\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}' = \boldsymbol{I} - \frac{\mathbf{1}\mathbf{1}'}{n}$$

and

$$\boldsymbol{M_1 y} = \Big[\boldsymbol{I} - \frac{\mathbf{1}\mathbf{1}'}{n}\Big]\boldsymbol{y} = \boldsymbol{y} - \bar{\boldsymbol{y}}.$$

It is easy to check that R^2 may be written as

$$\begin{aligned} R^2 &= 1 - \frac{\hat{\boldsymbol{u}}'\hat{\boldsymbol{u}}}{(\boldsymbol{y}-\bar{\boldsymbol{y}})'(\boldsymbol{y}-\bar{\boldsymbol{y}})} \\ &= \frac{(\hat{\boldsymbol{y}}-\bar{\boldsymbol{y}})'(\hat{\boldsymbol{y}}-\bar{\boldsymbol{y}})}{(\boldsymbol{y}-\bar{\boldsymbol{y}})'(\boldsymbol{y}-\bar{\boldsymbol{y}})} \\ &= \frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}, \end{aligned}$$

and that

$$0 \leq R^2 \leq 1.$$

In what follows we will call $||\boldsymbol{P_X y}||^2$ the *Regression Sum of Squares* (SSR), $||\boldsymbol{M_X y}||^2$ the *Error Sum of Squares* (SSE), and $||\boldsymbol{M_1 y}||^2$ the *Total Sum of Squares* (SST). With this terminology, R^2 may be written as

$$R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}}.$$

This popular regression statistic owes its name to the fact that it is equal to the square of the sample correlation between $\boldsymbol{y}$ and $\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$, i.e., letting

$$\hat{\rho} = \widehat{Corr}(\boldsymbol{y},\hat{\boldsymbol{y}}) = \frac{\widehat{Cov}(\boldsymbol{y},\hat{\boldsymbol{y}})}{\sqrt{\widehat{Var}(\boldsymbol{y})\widehat{Var}(\hat{\boldsymbol{y}})}},$$

we have $R^2 = \hat{\rho}^2$. It follows that in the simple $k = 1$ regression model, R^2 is equal to the square of the sample correlation coefficient between $\boldsymbol{y}$ and the single regressor $\boldsymbol{x}$.

5. Least Squares as a Conditional Expectation

Consider again the model

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{u}$$

and now assume that $E(\boldsymbol{u}|\boldsymbol{X}) = \boldsymbol{0}$ and that $V(\boldsymbol{u}|\boldsymbol{X}) = \sigma_u^2\boldsymbol{I}_n$. It follows directly that

$$E(\boldsymbol{y}|\boldsymbol{X}) = \boldsymbol{X\beta}, \quad \text{and} \quad V(\boldsymbol{y}|\boldsymbol{X}) = \sigma_u^2\boldsymbol{I}_n.$$

This is a *linear conditional expectation regression model* with a spherical (homoskedastic) error, of the form we have already encountered. As argued before, the OLS coefficients produce the optimal predictor of $E(\boldsymbol{y}|\boldsymbol{X})$ (see Theorem 3, Lecture 4). In fact, the assumption $E(\boldsymbol{u}|\boldsymbol{X}) = 0$ implies the uncorreletness condition $E(\boldsymbol{X'u}) = 0$, which in turn implies the "in-sample" orthogonality condition $\boldsymbol{X'u} = 0$, from which the OLS coefficients can be derived immediately as we have already seen. To see that $E(\boldsymbol{u}|\boldsymbol{X}) = 0$ implies $E(\boldsymbol{X}'\boldsymbol{u}) = 0$, observe that the first condition says that $\boldsymbol{u}$ is mean-independent of $\boldsymbol{X}$, while the second condition says that $\boldsymbol{u}$ and $\boldsymbol{X}$ are uncorrelated, and since mean-independence implies uncorrelateness, $E(\boldsymbol{u}|\boldsymbol{X}) = 0$ implies $E(\boldsymbol{X}'\boldsymbol{u}) = 0$.

The significance of interpreting least squares as a conditional expectation derives from that, in doing so, we are recasting the problem is *statistical terms.* As we have already seen the least squares solution has a rich *mathematical* (algebraic and geometric) structure, but the mathematical model alone cannot answer certain questions, like how "good" of an estimate of the true $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}}$? We thus presume that there is a *true parameter vector* $\boldsymbol{\beta}$, that $\hat{\boldsymbol{\beta}}$ is an *estimator* of it, and we are interested in the statistical properties of this estimator. This is, of course, the *statistical* point of view.

We make the following assumptions:

$$\begin{aligned}
(I) &\qquad \boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u} \\
(II) &\qquad E(\boldsymbol{u}|\boldsymbol{X}) = \boldsymbol{0} \\
(III) &\qquad E(\boldsymbol{u}'\boldsymbol{u}|\boldsymbol{X}) = \sigma_u^2 \boldsymbol{I}_n \\
(IV) &\qquad \boldsymbol{X}\ \textit{is a nonstochastic matrix of rank}\, k \\
(V) &\qquad \boldsymbol{u} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n).
\end{aligned}$$

Assumption (I) says that the function relating $\boldsymbol{y}$ to $\boldsymbol{X}$ is linear and that $\boldsymbol{X}$ contains exactly the necessary variables to explain $\boldsymbol{y}$, without excluding any relevant or including any irrelevant ones. Assumptions (II) and (III) say that the residual error, once the effect of the regressors on the dependent variables has been accounted for, has mean zero and constant variance $\sigma^2 \boldsymbol{I}_n$. Assumption (IV) says that the $n \times k$ regressor matrix $\boldsymbol{X}$ is fixed, i.e., non-random (for example, the design matrix of an experiment) and has full rank k, so that $(\boldsymbol{X}'\boldsymbol{X})$ is invertible. The full rank condition means that none of the regressors is a linear function of the other regressors. Assumptions (I)-(IV) are called *weak.* Finally, assumption (IV) is a *strong* assumption that specifies the entire distribution and not just the mean and variance of the error term $\boldsymbol{u}$. It says that $\boldsymbol{u}$ it is normally distributed with mean and variance as in (II) and (III). This assumption is necessary in order to be able to do inference in small samples, but as the sample size n grows, it can be dropped on account of the Central Limit Theorem (CLT) that makes this assumption unnecessary, at least asymptotically.

We say that $\hat{\boldsymbol{\theta}}$ is an unbiased estimator of the parameter $\boldsymbol{\theta}$, if $E(\hat{\boldsymbol{\theta}}) = \boldsymbol{\theta}$. That, under our assumptions, $\hat{\boldsymbol{\beta}}$ is unbiased for $\boldsymbol{\beta}$ follows immediately,

$$\begin{aligned} E(\hat{\boldsymbol{\beta}}|\boldsymbol{X}) &= E[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}|\boldsymbol{X}] \\ &= E[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{u})|\boldsymbol{X}] \\ &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'E(\boldsymbol{u}|\boldsymbol{X}) \\ &= \boldsymbol{\beta}. \end{aligned}$$

The variance of $\hat{\boldsymbol{\beta}}$ is also easy to derive,

$$\begin{aligned} V(\hat{\boldsymbol{\beta}}|\boldsymbol{X}) &= E[(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})'|\boldsymbol{X}] \\ &= E[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{u}\boldsymbol{u}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}|\boldsymbol{X}] \\ &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'E[\boldsymbol{u}\boldsymbol{u}'|\boldsymbol{X}]\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} \\ &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\sigma_u^2\boldsymbol{I}_n\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} \\ &= \sigma_u^2(\boldsymbol{X}'\boldsymbol{X})^{-1}. \end{aligned}$$

An estimator of the error variance σ_u^2 may be obtained by the sample variance of the regression residuals $\hat{u}$,

$$\hat{\sigma}_u^2 = \frac{\hat{\boldsymbol{u}}'\hat{\boldsymbol{u}}}{n} = \frac{1}{n}\sum_{i=1}^n \hat{u}_i^2.$$

But this a biased estimator of σ_u^2, since

$$\begin{aligned} E(n\hat{\sigma}_u^2|\boldsymbol{X}) &= E(\hat{\boldsymbol{u}}'\hat{\boldsymbol{u}}|\boldsymbol{X}) \\ &= E(\boldsymbol{u}'\boldsymbol{M_X}\boldsymbol{u}|\boldsymbol{X}) \\ &= E(\text{trace}(\boldsymbol{u}'\boldsymbol{M_X}\boldsymbol{u})|\boldsymbol{X}) && \text{[because } \boldsymbol{u}'\boldsymbol{M_X}\boldsymbol{u} \text{ is a scalar]} \\ &= E(\text{trace}(\boldsymbol{M_X}\boldsymbol{u}\boldsymbol{u}')|\boldsymbol{X}) && \text{[by a property of the trace]} \\ &= \text{trace}[E(\boldsymbol{M_X}\boldsymbol{u}\boldsymbol{u}'|\boldsymbol{X})] && \text{[because the trace is a linear operator]} \\ &= \text{trace}[\boldsymbol{M_X}E(\boldsymbol{u}\boldsymbol{u}'|\boldsymbol{X})] \\ &= \text{trace}[\boldsymbol{M_X}\sigma_u^2\boldsymbol{I}_n] \\ &= \sigma_u^2(n-k). && \text{[because trace } \boldsymbol{M_X} = \text{rank } \boldsymbol{M_X} = n-k\text{].} \end{aligned}$$

Therefore,

$$E(\hat{\sigma}_u^2) = \Big(\frac{n-k}{n}\Big)\sigma_u^2$$

and we see that $\hat{\sigma}_u^2$ is a biased estimator of σ_u^2. An unbiased estimator is thus given by

$$s_u^2 = \frac{\hat{\boldsymbol{u}}'\hat{\boldsymbol{u}}}{n-k} = \frac{1}{n-k}\sum_{i=1}^n \hat{u}_i^2,$$

where we divide the sum of squared residuals not by the sample size n, but by the *degrees of freedom* $n-k$.

We are now in a position to state a well-known optimality result for OLS, the so called *Gauss-Markov theorem.*

Theorem 2. (Gauss-Markov). *Under Assumption (I)-(IV), the OLS estimator $\hat{\boldsymbol{\beta}}$ is the Best Linear (in $\boldsymbol{y}$) Unbiased Estimator (BLUE) of $\boldsymbol{\beta}$, i.e., if $\tilde{\boldsymbol{\beta}}$ is another linear unbiased estimator of $\boldsymbol{\beta}$, then $V(\tilde{\boldsymbol{\beta}})-V(\hat{\boldsymbol{\beta}})$ is a positive semidefinite matrix.*

Proof. Since $\tilde{\boldsymbol{\beta}}$ is a linear function of $\boldsymbol{y}$, we can write as

$$\tilde{\boldsymbol{\beta}} = \boldsymbol{A}\boldsymbol{y} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} + \boldsymbol{C}\boldsymbol{y}$$

where $\boldsymbol{C} = \boldsymbol{A} - (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$. Substituting $\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{u}$ for $\boldsymbol{y}$ we obtain

$$\begin{aligned}\tilde{\boldsymbol{\beta}} &= [(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' + \boldsymbol{C}](\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{u}) \\ &= \boldsymbol{\beta} + \boldsymbol{C}\boldsymbol{X}\boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{u} + \boldsymbol{C}\boldsymbol{u}.\end{aligned}$$

For $\tilde{\boldsymbol{\beta}}$ to be unbiased, we must have $\boldsymbol{C}\boldsymbol{X}=\boldsymbol{0}$. Imposing this condition yields

$$\tilde{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{u} + \boldsymbol{C}\boldsymbol{u}.$$

The variance of $\tilde{\boldsymbol{\beta}}$ is now given by

$$\begin{aligned}V(\tilde{\boldsymbol{\beta}}|\boldsymbol{X}) &= E[(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta})(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta})'|\boldsymbol{X}] \\ &= E\Big[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{u}\boldsymbol{u}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{u}\boldsymbol{u}'\boldsymbol{C}' \\ &\qquad + \boldsymbol{C}\boldsymbol{u}\boldsymbol{u}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} + \boldsymbol{C}\boldsymbol{u}\boldsymbol{u}'\boldsymbol{C}'\Big|\boldsymbol{X}\Big] \\ &= \sigma_u^2\Big\{(\boldsymbol{X}'\boldsymbol{X})^{-1} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{C}' + \boldsymbol{C}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} + \boldsymbol{C}\boldsymbol{C}'\Big\}.\end{aligned}$$

Imposing again the condition $\boldsymbol{C}\boldsymbol{X}=\boldsymbol{0}$, we obtain

$$V(\tilde{\boldsymbol{\beta}}|\boldsymbol{X}) = \sigma_u^2\Big\{(\boldsymbol{X}'\boldsymbol{X})^{-1} + \boldsymbol{C}\boldsymbol{C}'\Big\}.$$

Thus,

$$V(\hat{\boldsymbol{\beta}}|\boldsymbol{X}) - V(\tilde{\boldsymbol{\beta}}|\boldsymbol{X}) = \sigma_u^2\boldsymbol{C}\boldsymbol{C}',$$

which is a positive semidefinite matrix. □

It is worth noting that the spherical assumption $V(\boldsymbol{u}|\boldsymbol{X}) = \sigma_u^2\boldsymbol{I}_n$ is crucial for the validity of the Gauss-Markov theorem. If we instead had a *non-spherical* (heteroskedastic and/or autocorrelated) error with $V(\boldsymbol{u}|\boldsymbol{X}) = \boldsymbol{\Omega}$, the variance of $\tilde{\boldsymbol{\beta}}$ would be

$$(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Omega}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Omega}\boldsymbol{C}' + \boldsymbol{C}\boldsymbol{\Omega}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} + \boldsymbol{C}\boldsymbol{\Omega}\boldsymbol{C}',$$

and we would not be able to draw any conclusion about the relative efficiency of $\hat{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\beta}}$. We will see later that in this non-spherical error case, another estimator is BLUE, namely the Generalized Least Squares (GLS) estimator.

This often quoted optimality property of the OLS estimator, namely that it is BLUE, should not, however, be overvalued. It only says that OLS is best (has lowest variance) among a very restricted class of estimators, namely estimators that are linear(!) in $\boldsymbol{y}$. There is really no reason to impose the linearity condition other than to make the OLS estimator "optimal". As we shall see later, relaxing this arbitrary and unnecessary linearity assumption makes OLS just another estimator among many. The LAD estimator, for example, that we will also study later, is not linear in $\boldsymbol{y}$, but may be shown to be unbiased and have superior performance (smaller variance) than the OLS estimator under error distributions with thick tails. Therefore, we should always remember that there is no BUE, and that the Gauss-Markov Theorem only says that OLS is BLUE, not BUE!

6. Residual Regression and the Frisch-Waugh-Lovell Theorem.

Partition $\boldsymbol{X} = (\boldsymbol{X_1},\ \boldsymbol{X_2})$ and

$$\boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta_1} \\ \boldsymbol{\beta_2} \end{pmatrix},$$

and rewrite the regression model as

$$\boldsymbol{y} = \boldsymbol{X_1}\boldsymbol{\beta_1} + \boldsymbol{X_2}\boldsymbol{\beta_2} + \boldsymbol{u}.$$

Observe that the OLS estimator $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_{\boldsymbol{1}}{}', \hat{\boldsymbol{\beta}}_{\boldsymbol{2}}{}')'$ can be obtained by regressing $\boldsymbol{y}$ on $\boldsymbol{X} = (\boldsymbol{X_1},\ \ \boldsymbol{X_2})$, so we can write

$$y = \boldsymbol{X_1}\hat{\boldsymbol{\beta}}_{\boldsymbol{1}} + \boldsymbol{X_2}\hat{\boldsymbol{\beta}}_{\boldsymbol{2}} + \hat{\boldsymbol{u}}.$$

The OLS coefficient vector $\hat{\boldsymbol{\beta}}$ can be written as

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = \begin{pmatrix} \boldsymbol{X_1}'\boldsymbol{X_1} & \boldsymbol{X_1}'\boldsymbol{X_2} \\ \boldsymbol{X_2}'\boldsymbol{X_1} & \boldsymbol{X_2}'\boldsymbol{X_2} \end{pmatrix}^{-1} \begin{pmatrix} \boldsymbol{X_1}'y \\ \boldsymbol{X_2}'\boldsymbol{y} \end{pmatrix}.$$

To proceed we need the following lemma on partition matrices.

Lemma 1. (Inverse of Partition Matrices). *Suppose that a positive definite matrix $\boldsymbol{Q}$ is partitioned as*

$$\boldsymbol{Q} = \begin{pmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{pmatrix},$$

where the diagonal blocks $\boldsymbol{A}$ and $\boldsymbol{D}$ are square matrices. Then

$$\boldsymbol{Q}^{-1} = \begin{pmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{pmatrix}^{-1} = \begin{pmatrix} \boldsymbol{E}^{-1} & -\boldsymbol{E}^{-1}\boldsymbol{B}\boldsymbol{D}^{-1} \\ -\boldsymbol{D}^{-1}\boldsymbol{C}\boldsymbol{E}^{-1} & \boldsymbol{F}^{-1} \end{pmatrix},$$

where

$$\begin{aligned} \boldsymbol{E}^{-1} &= \boldsymbol{A}^{-1} + \boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{F}^{-1}\boldsymbol{C}\boldsymbol{A}^{-1} &&= (\boldsymbol{A} - \boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C})^{-1}, \\ \boldsymbol{F}^{-1} &= \boldsymbol{D}^{-1} + \boldsymbol{D}^{-1}\boldsymbol{C}\boldsymbol{E}^{-1}\boldsymbol{B}\boldsymbol{D}^{-1} &&= (\boldsymbol{D} - \boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B})^{-1}. \end{aligned}$$

Using Lemma 1, we have

$$\begin{aligned} &\begin{pmatrix} \boldsymbol{X_1}'\boldsymbol{X_1} & \boldsymbol{X_1}'\boldsymbol{X_2} \\ \boldsymbol{X_2}'\boldsymbol{X_1} & \boldsymbol{X_2}'\boldsymbol{X_2} \end{pmatrix}^{-1} \\ &= \begin{pmatrix} (\boldsymbol{X_1}'\boldsymbol{M_2}\boldsymbol{X_1})^{-1} & -(\boldsymbol{X_1}'\boldsymbol{M_2}\boldsymbol{X_1})^{-1}\boldsymbol{X_1}'\boldsymbol{X_2}(\boldsymbol{X_2}'\boldsymbol{X_2})^{-1} \\ -(\boldsymbol{X_1}'\boldsymbol{M_2}\boldsymbol{X_1})^{-1}\boldsymbol{X_1}'\boldsymbol{X_2}(\boldsymbol{X_2}'\boldsymbol{X_2})^{-1} & (\boldsymbol{X_2}'\boldsymbol{M_1}\boldsymbol{X_2})^{-1} \end{pmatrix} \end{aligned}$$

where

$$\begin{aligned} M_1 &= I_n - X_1(X_1'X_1)^{-1}X_1' \\ M_2 &= I_n - X_2(X_2'X_2)^{-1}X_2'. \end{aligned}$$

Thus,

$$\begin{aligned} \begin{pmatrix} \hat\beta_1 \\ \hat\beta_2 \end{pmatrix} &= \begin{pmatrix} X_1'X_1 & X_1'X_2 \\ X_2'X_1 & X_2'X_2 \end{pmatrix}^{-1} \begin{pmatrix} X_1'y \\ X_2'y \end{pmatrix} \\ &= \begin{pmatrix} (X_1'M_2X_1)^{-1} & -(X_1'M_2X_1)^{-1}X_1'X_2(X_2'X_2)^{-1} \\ -(X_1'M_2X_1)^{-1}X_1'X_2(X_2'X_2)^{-1} & (X_2'M_1X_2)^{-1} \end{pmatrix} \begin{pmatrix} X_1'y \\ X_2'y \end{pmatrix} \\ &= \begin{pmatrix} (X_1'M_2X_1)^{-1}X_1'M_2y \\ (X_2'M_1X_2)^{-1}X_2'M_1y \end{pmatrix}. \end{aligned}$$

We have proven the following "lovel(l)y" theorem.

Theorem 3. (Frisch-Waugh-Lovell Theorem). *In the linear regression model*

$$\boldsymbol{y} = \boldsymbol{X_1}\boldsymbol{\beta_1} + \boldsymbol{X_2}\boldsymbol{\beta_2} + \boldsymbol{u},$$

the OLS estimator for $\boldsymbol{\beta_1}$ and the OLS residuals $\hat{\boldsymbol{u}}$ may be computed by either the full OLS regression of $\boldsymbol{y}$ on $\boldsymbol{X} = (\boldsymbol{X_1}, \boldsymbol{X_2})$ or via the following procedure:

(i) Regress $\boldsymbol{y}$ on $\boldsymbol{X_2}$ and obtain residuals $\tilde{\boldsymbol{y}}_{\mathbf{1}} = \boldsymbol{M_2 y}$;

(ii) Regress $\boldsymbol{X_1}$ on $\boldsymbol{X_2}$ and obtain residuals $\tilde{\boldsymbol{X}}_{\mathbf{1}} = \boldsymbol{M_2 X_1}$;

(iii) Regress $\tilde{\boldsymbol{y}}_1$ on $\tilde{\boldsymbol{X}}_{\mathbf{1}}$ to obtain $\hat{\boldsymbol{\beta}}_{\mathbf{1}}$ and residuals $\hat{\boldsymbol{u}}$, i.e.,

$$\hat{\boldsymbol{\beta}}_{\mathbf{1}} = (\boldsymbol{X_1}'\boldsymbol{M_2 X_1})^{-1}\boldsymbol{X}_1'\boldsymbol{M_2 y}$$

and

$$\hat{\boldsymbol{u}} = \boldsymbol{M_2 y} - \boldsymbol{M_2 X_1}(\boldsymbol{X_1}'\boldsymbol{M_2 X_1})^{-1}\boldsymbol{M_2 y X_1}'\boldsymbol{M_2 y}.$$

Also, a completely analogous procedure yields

$$\hat{\boldsymbol{\beta}}_{\mathbf{2}} = (\boldsymbol{X_2}'\boldsymbol{M_1 X_2})^{-1}\boldsymbol{X_2}'\boldsymbol{M_1 y}.$$

This theorem was initially advocated as a computational device that can break up large least squares problems into smaller ones, but today its value is mainly theoretical. There is, however, a nice application of this theorem in ploting $\boldsymbol{y}$ against a multivariate $\boldsymbol{X}$. Such plots give valuable information about the functional relation between y and the $x's$ and the potential adequacy of a linear fit. When X is multidimensional, however, such plots become impossible. Using Theorem 2, we can isolate the influence of each of the regressors on y. Assume, for example, that there are 3 regressors (excluding the intercept). Then, we can (a) regress y on x_2 and x_3 (without an intercept) and obtain residuals $\tilde{y}_1$, (b) regress x_1 on x_2 and x_3 (without an intercept) and obtain residuals $\tilde{x}_1$, and (c) plot $\tilde{y}_1$ against $\tilde{x}_1$. Repeating for the other 2 regressors, we obtain a set of 3 plots, one for each regressor, called *partial residual plots.* However, although useful, such plots should be used with caution since for constructing the plot of $\tilde{y}_1$ against $\tilde{x}_1$ for example, we used the assumption that x_2 and x_3 enter linearly, which may not be true.

7. RESTRICTED LEAST SQUARES.

Consider the *restricted* regression model

$$\begin{aligned} &\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{u} \\ \text{such that} \qquad &\boldsymbol{R\beta} = \boldsymbol{r} \end{aligned}$$

where $\boldsymbol{R}$ is $q \times k$ matrix and $\boldsymbol{r}$ is a $q \times 1$ vector. The Lagrangian of the problem is given by

$$S(\boldsymbol{\beta}) = (\boldsymbol{y} - \boldsymbol{X\beta})'(\boldsymbol{y} - \boldsymbol{X\beta}) - 2\boldsymbol{\lambda}(\boldsymbol{R\beta} - \boldsymbol{r})$$

where $\boldsymbol{\lambda}$ is a $q \times 1$ vector of Langrange multipliers. The F.O.C. yield

$$\begin{aligned}\frac{\partial S(\hat{\boldsymbol{\beta}}_R)}{\partial \boldsymbol{\beta}} &= -2\boldsymbol{X}'\boldsymbol{y} + 2(\boldsymbol{X}'\boldsymbol{X})\hat{\boldsymbol{\beta}}_R - 2\boldsymbol{R}'\hat{\boldsymbol{\lambda}} = \mathbf{0} \\ \frac{\partial S(\hat{\boldsymbol{\beta}}_R)}{\partial \boldsymbol{\lambda}} &= -2(\boldsymbol{R}\hat{\boldsymbol{\beta}}_R - \boldsymbol{r}) = \mathbf{0}.\end{aligned}$$

The first equation yields

$$\begin{aligned}\hat{\boldsymbol{\beta}}_R &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'\hat{\boldsymbol{\lambda}} \\ &= \hat{\boldsymbol{\beta}} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'\hat{\boldsymbol{\lambda}}.\end{aligned}$$

Premultiplying with $\boldsymbol{R}$, we obtain

$$\boldsymbol{R}\hat{\boldsymbol{\beta}}_R = \boldsymbol{R}\hat{\boldsymbol{\beta}} + \boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'\hat{\boldsymbol{\lambda}}.$$

By the restriction, the left hand side of this equation is equal to $\boldsymbol{r}$, and solving for $\hat{\boldsymbol{\lambda}}$ we obtain

$$\hat{\boldsymbol{\lambda}} = \Big[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'\Big]^{-1}(\boldsymbol{r} - \boldsymbol{R}\hat{\boldsymbol{\beta}}).$$

Substituting into the $\hat{\boldsymbol{\beta}}_R$ equation above we finally obtain

$$\hat{\boldsymbol{\beta}}_R = \hat{\boldsymbol{\beta}} - (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'\Big[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'\Big]^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r}).$$

If, by some miracle, the unrestricted OLS estimates $\hat{\boldsymbol{\beta}}$ should happen to satisfy the restriction *exactly in the sample*, i.e., if $\boldsymbol{R}\hat{\boldsymbol{\beta}} = \boldsymbol{r}$, the second term in the above expression vanishes and $\hat{\boldsymbol{\beta}}_R = \hat{\boldsymbol{\beta}}$. More realistically, when the discrepancy between $\boldsymbol{R}\hat{\boldsymbol{\beta}}$ and $\boldsymbol{r}$ is small (i.e., when the data more or less agree with the restriction), the two estimates $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}_R$ will be close, but when the discrepancy between $\boldsymbol{R}\hat{\boldsymbol{\beta}}$ and $\boldsymbol{r}$ is large (i.e., when the restriction puts a strain on the data), $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}_R$ will deviate considerably.

8. The Neoclassical Regression Model.

Consider again the linear regression model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}$$

and recall the assumptions we have made so far,

(i) $\boldsymbol{X}$ is of full rank, i.e. $\text{rank}(\boldsymbol{X}) = k$;

(ii) $E(\boldsymbol{u}|\boldsymbol{X}) = \mathbf{0}$; and

(iii) $V(\boldsymbol{u}|\boldsymbol{X}) = \sigma_u^2 I$.

Under these assumptions we were able to show that the OLS coefficients $\hat{\beta}$ are unbiased and BLUE. But we are not content yet. We would like to be able to construct confidence intervals for the population parameter $\boldsymbol{\beta}$, and we would also like to be able to test hypotheses regarding β. This would require knowledge of the *distribution* of $\hat{\beta}$, but all we have gotten so far out of our assumptions are the mean and variance of this distribution. We clearly need some extra structure. In the future, we will explore the behavior of $\hat{\beta}$ as the sample size n grows large, and we will show that the current assumptions will be enough to determine the *asymptotic* distribution of $\hat{\beta}$ under random sampling. For now, however, we will assume that n is fixed and small, and impose a new assumption that will determine the distribution of $\hat{\beta}$ at any sample size:

$$\text{(iv)}\quad \boldsymbol{u}|\boldsymbol{X} \sim N(\mathbf{0}, \sigma_u^2 \boldsymbol{I}_k).$$

This new assumption is very strong: it says that, not only is the mean and variance of $u|X$ are as given above, but also the conditional distribution of u given X is *normal*! This model with the extra assumption (iv) is called the *neoclassical regression model.*

Assumption (iv) yields immediately that $\boldsymbol{y}|\boldsymbol{X} \sim N(\boldsymbol{X\beta}, \sigma_u^2 \boldsymbol{I}_k)$, and since the OLS vector $\hat{\beta}$ is a linear function of y we have

$$\hat{\beta}|X \sim N(\beta, \sigma_u^2 (X'X)^{-1}).$$

This is really nice, the only difficulty being that σ_u^2 is unknown. We could replace it, of course, with $\hat{s}_u^2$ but as we will see presently, this would change the (small sample) distribution of $\hat{\beta}$.

9. NORMAL AND RELATED DISTRIBUTIONS.

9.1. THE MULTIVARIATE NORMAL DISTRIBUTION - A ROMANCE OF MANY DIMENSIONS.

For an amusing account of how $\mathbb{R}^3$ would be perceived by creatures living in $\mathbb{R}^2$, read *Flatland: A Romance of Many Dimensions*, by Edwin A. Abbott. This novel, published in 1884, may help you imagine a physical space of four or more dimensions.

The density of a $(k \times 1)$, $MVN_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ random vector $\boldsymbol{X}$ is given by

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^{k/2} |\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{(\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}{2} \right).$$

The bivariate normal $BVN(\mu_1, \mu_2, \rho, \sigma_1, \sigma_2)$ is given by

$$f(x_1, x_2) = \frac{1}{2\pi\sqrt{\sigma_1^2 \sigma_2^2 (1-\rho^2)}} \exp\left(\frac{-z_1^2 + 2\rho z_1 z_2 - z_2^2}{2(1-\rho^2)} \right).$$

F 3,
fe

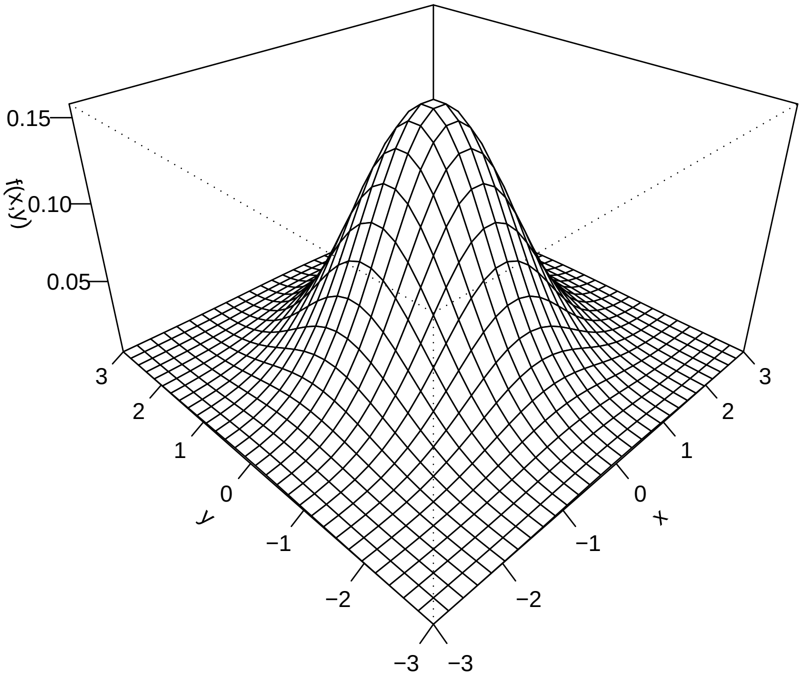

FIGURE 7. The standard bivariate normal distribution.

where $z_1 = (x_1 - \mu_1)/\sigma_1$ and $z_2 = (x_2 - \mu_2)/\sigma_2$. The following lemma describes how to construct samples from a bivariate normal distribution.

LEMMA 2. *Let U_1, U_2 be independent $N(0,1)$ variables, and let ρ be any constant such that $|\rho| < 1$. Then the random variables*

$$Z_1 = U_1, \qquad Z_2 = \rho U_1 + \sqrt{(1-\rho^2)}U_2$$

follow a standard joint bivariate normal distribution with correlation ρ, i.e.,

$$(Z_1, Z_2) \sim BVN\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}\right).$$

Furthermore, for constants $\mu_1, \mu_2, \sigma_1 > 0, \sigma_2 > 0$, the variables

$$X_1 = \mu_1 + \sigma_1 Z_1, \qquad X_2 = \mu_2 + \sigma_2 Z_2$$

follow a joint bivariate normal distribution, i.e., $(X_1, X_2) \sim BVN(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu} = (\mu_1, \mu_2)'$ and

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}.$$

It is important to note here that asserting *joint* normality of X and Y is a much stronger statement than claiming that X and Y are *marginally* normal. As the following examples show, two variables may be marginally normal but jointly non-normal.

Example 1. To see that joint normality implies marginal normality, but the reverse is not generally true, take $X \sim N(0,1)$, and recall that, by the probability transformation, $\Phi(X) \sim U[0,1]$. Now define the transformation T by

$$T(x) = \begin{cases} x, & 0 \le x \le \frac{1}{2} \\ \frac{3}{2} - x, & \frac{1}{2} < x \le 1. \end{cases}$$

Then $V = T(U)$ is $U[0,1]$ and therefore $Y = \Phi^{-1}(V) = \Phi^{-1}(T(\Phi(X))$ is also $N(0,1)$ (note that this construction guarantees that X and Y are not independent). Both X and Y are *marginally* standard normal, but there are many ways to argue that they are not *jointly* normal. For example, the variable $W = (X+Y)/\sqrt{2}$ cannot be standard normal (as it should be if X and Y were jointly normal) because it can not achieve some positive values, as a standard normal should be able to. ■

Example 2. Consider the joint pdf

$$f(x,y) = 2I\{xy > 0\}\phi(x)\phi(y),$$

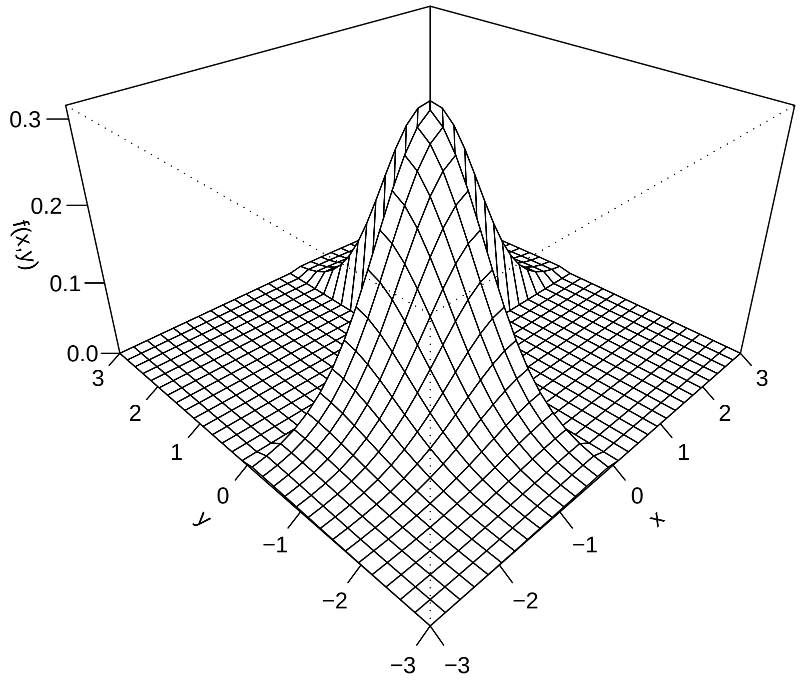

FIGURE 8. A non-normal bivariate distribution with standard normal marginals.

where $\phi()$ is the standard normal pdf (see Figure 8). The marginal pdf of X is

$$f_x(x) = \int_{-\infty}^{\infty} f(x,y)dy = 2\phi(x)\int_{-\infty}^{\infty} I\{xy > 0\}\phi(y)dy.$$

For $x > 0$, $I\{xy > 0\}\phi(y) = \phi(y)$ for $y > 0$, and $I\{xy > 0\}\phi(y) = 0$ for $y \leq 0$. So for $x > 0$,

$$\int_{-\infty}^{\infty} I\{xy > 0\}\phi(y)dy = \int_{0}^{\infty} \phi(y)dy = \tfrac{1}{2}.$$

Similarly for $x \leq 0$. So $f_x(x) = 2\phi(x)\frac{1}{2} = \phi(x)$, that is $X \sim N(0,1)$. By symmetry $Y \sim N(0,1)$ also, so both marginals are standard normal. But the joint distribution is not bivariate normal. Also, the CEF is not linear and the conditional pdf's are not normal either (Show this as an exercise). ■

The following important theorem characterizes multivariate normality.

THEOREM 4. *Let $\boldsymbol{X}$ be a k-dimensional random vector. Then $\boldsymbol{X} \sim MVN_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ if and only if for each $\boldsymbol{t} \in \mathbb{R}^k$, $\boldsymbol{t'X} \sim N(\boldsymbol{t'\mu}, \boldsymbol{t'\Sigma t})$.*

In words, the theorem says that a random vector is jointly normal if and only if *any* linear combination of its components is also normal.

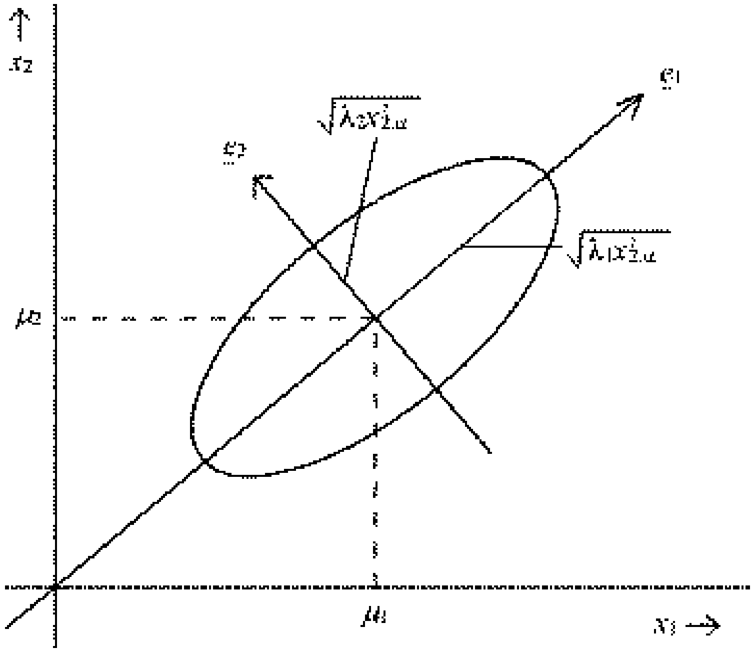
x_2
e_1
$\sqrt{\lambda_2\chi^2_{2,\alpha}}$
e_2
$\sqrt{\lambda_1\chi^2_{2,\alpha}}$
μ_2
μ_1
x_1

FIGURE 9. The geometry of the Multivariate Normal Distribution.

THEOREM 5. (CRAMER'S THEOREM) *If the sum of two independent random variables is normally distributed, then each of the summands is also normally distributed.*

9.2. THE GEOMETRY OF THE MULTIVARIATE NORMAL DISTRIBUTION.

The geometry of the multivariate normal distribution can be investigated by considering the orientation and shape of the prediction ellipse as depicted in the following diagram:

The $(1 - a) \times 100\%$ prediction ellipse above is centered on the population means μ_1 and μ_2. The ellipse has axes pointing in the directions of the eigenvectors $\boldsymbol{e}_1, \boldsymbol{e}_2, ..., \boldsymbol{e}_p$. Here, in this diagram for the bivariate normal, the longest axis of the ellipse points in the direction of the first eigenvector $\boldsymbol{e}_1$ and the shorter axis is perpendicular to the first, pointing in the direction of the second eigenvector $\boldsymbol{e}_2$. The corresponding half-lengths of the axes are

$$l_j = \sqrt{\lambda_j \chi^2_{p,\alpha}}$$

The plot above shows the lengths of these axes within the ellipse.

The volume (or area if $p = 1$) of the hyper-ellipse is equal to

$$\frac{2\pi^{p/2}}{p\Gamma\left(\frac{p}{2}\right)}(\chi^2_{p,\alpha})^{p/2}|\boldsymbol{\Sigma}|^{1/2}.$$

Note that the area is proportional to the square-root of the *generalized variance* given by the square root of the determinant of the variance-covariance matrix, $|\boldsymbol{\Sigma}|^{1/2}$. To compute the gamma function $\Gamma(x)$ in this expression note that, for p even,

$$\Gamma\left(\frac{p}{2}\right) = \left(\frac{p}{2} - 1\right)!$$

while for p odd,

$$\Gamma\left(\frac{p}{2}\right) = \frac{1 \times 3 \times 5 \times \cdots \times (p-2) \times \sqrt{\pi}}{2^{(p-1)/2}}.$$

9.3. The Chi-Square Distribution.

The pdf of a *non-central chi-square* random variable X with $n > 0$ *degrees of freedom* and *non-centrality parameter* $\delta \geq 0$, denoted by $\chi^2_n(\delta)$, is given by

$$f(y;\delta) = \frac{\exp[-\frac{1}{2}(y+\delta)]}{2^{n/2}} \sum_{j=0}^{\infty} \frac{y^{n/2+j-1}\delta^j}{\Gamma(n/2+j)2^{2j}j!}, \qquad y \geq 0.$$

For $\delta = 0$ we obtain the pdf of the *central chi-square distribution*, denoted by χ^2_n,

$$f(y) = \frac{y^{\frac{n}{2}-1}e^{-\frac{y}{2}}}{\Gamma(\frac{n}{2})\,2^{\frac{n}{2}}}, \qquad y \geq 0.$$

Lemma 3. *If $X \sim MVN_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then*

$$\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X} \sim \chi^2_k(\delta)$$

with noncentrality parameter $\delta = \boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$. Also,

$$(X-\mu)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}-\boldsymbol{\mu}) \sim \chi^2_k,$$

i.e., a central chi-squared distribution with k degrees of freedom.

9.4. THE F DISTRIBUTION.

The positive random variable Y with pdf

$$f(y) = \frac{\Gamma[\frac{1}{2}(n+k)](n/k)^{n/2}x^{(n-2)/2}}{\Gamma(n/2)\Gamma(k/2)[1+(n/k)x]^{(n+k)/2}}, \qquad y \geq 0,$$

is said to follow the F distribution with degrees of freedom.

LEMMA 4. *If $Y_1 \sim \chi^2_{n_1}$, $Y_2 \sim \chi^2_{n_2}$, and Y_1 and Y_2 are independent, then*

$$F = \frac{Y_1/n_1}{Y_2/n_2} \sim F_{n_1,n_2}.$$

9.5. THE STUDENT'S T DISTRIBUTION.

The random variable X with pdf

$$f(x) = \frac{\Gamma[(n+1)/2]}{(\pi n)^{1/2}\Gamma(n/2)[1+(x^2/n)]^{(n+1)/2}}, \quad x \in \mathbb{R}$$

where n is a positive integer, is said to follow the t_n distribution.

LEMMA 5. *If $Z \sim N(0,1)$, and $Y \sim \chi^2_n$, independent then,*

$$T = \frac{Z}{\sqrt{Y/n}} \sim t_n.$$

Observe that

$$T^2 = \frac{Z^2/1}{Y/n} = F_{1,n}.$$

The t_1 distribution is the Cauchy. The t_n distribution has $n-1$ moments.

10. CONFIDENCE REGIONS.

Let $\boldsymbol{R}$ be a $q \times k$ matrix and consider the $q \times 1$ vectors $\boldsymbol{R\beta}$ and $\boldsymbol{R\hat{\beta}}$ of population restrictions and sample estimates. Using the normality of $\hat{\boldsymbol{\beta}}$ we obtain

$$\boldsymbol{R\hat{\beta}} \sim N(\boldsymbol{R\beta}, \sigma_u^2 \boldsymbol{R}(\boldsymbol{X'X})^{-1}\boldsymbol{R'}).$$

Therefore, the centered and normalized quadratic form

$$(\boldsymbol{R\hat{\beta}} - \boldsymbol{R\beta})'[\sigma_u^2 \boldsymbol{R}(X'X)^{-1}\boldsymbol{R'}]^{-1}(\boldsymbol{R\hat{\beta}} - \boldsymbol{R\beta}) \sim \chi^2_q,$$

is a χ^2 random variable with q degrees of freedom. It also follows that, if σ_u^2 is known, a $100(1-\alpha)\%$ *confidence region* (CR) for $\boldsymbol{R\beta}$ is given by

$$(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{R\beta})'[\sigma_u^2 \boldsymbol{R}(X'X)^{-1}\boldsymbol{R}']^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{R\beta}) < \chi^2_{q,\alpha}$$

where $\chi^2_{q,\alpha}$ is the number for which $\Pr(\chi^2_q < \chi^2_{q,\alpha}) = 1-\alpha$.

In actual application, however, σ_u^2 is unknown and has to be replaced by $\hat{s}_u^2$. Since

$$\frac{\hat{\boldsymbol{u}}'\hat{\boldsymbol{u}}}{\sigma_u^2} = \frac{(n-k)\hat{s}_u^2}{\sigma_u^2} \sim \chi^2_{n-k},$$

it follows that

$$\frac{(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{R\beta})'[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}']^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{R\beta})/q}{\hat{\boldsymbol{u}}'\hat{\boldsymbol{u}}/(n-k)} \sim F_{q,n-k}$$

since the ratio of two χ^2 random variables normalized by their respective degrees of freedom is distributed as F. Thus, in the case where σ_u^2 is unknown, a $100(1-\alpha)\%$ confidence region for $\boldsymbol{R\beta}$ is given by

$$\frac{(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{R\beta})'[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}']^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{R\beta})/q}{\hat{\boldsymbol{u}}'\hat{\boldsymbol{u}}/(n-k)} \leq F_{q,n-k,\alpha} \qquad (*)$$

where $F_{q,n-k,\alpha}$ is the number for which $\Pr(F_{q,n-k} \leq F_{q,n-k,\alpha}) = 1-\alpha$.

Various special cases now follow immediately. If $\boldsymbol{R}$ is $1 \times k$, i.e. $q = 1$, then $\theta = \boldsymbol{R\beta}$ is a scalar, and the $100 \times (1-\alpha)\%$ confidence region reduces to a $100 \times (1-\alpha)\%$ confidence *interval* (CI) given by

$$\frac{(\hat{\theta}-\theta)^2}{\hat{\sigma}_\theta^2} \sim F_{1,n-k,\alpha}$$

where $\hat{\sigma}^2_\theta = \hat{s}_u^2[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}']^{-1}$. From Lemma 5, we have that $F_{1,n} = t_n^2$, so

$$\frac{\hat{\theta}-\theta}{\sigma_{\hat{\boldsymbol{\theta}}}} \sim t_{n-k,\alpha}$$

which yields the familiar $100 \times (1-\alpha)\%$ CI for θ:

$$\hat{\theta} - t_{n-k,\alpha/2}\ \hat{\sigma}_\theta < \theta < \hat{\theta} + t_{n-k,\alpha/2}\ \hat{\sigma}_\theta.$$

An 95%, say, confidence interval for a parameter θ can be likened to a person that tells the truth 95% of the time, but we do not know whether a particular statement he makes is true or not. Likewise, an 95% confidence interval is calculated such that it includes the true value of the estimated parameter 95% of the time. We do not know, however, if the interval we have is one of those that are correct or not. Put differently, if the random experiment is repeated 100 times, in 95 of them our CI will contain the true parameter θ_0, and in 5 of them it will not.

11. HYPOTHESIS TESTING.

Now assume that we wish to test the hypothesis

$$H_0 : \boldsymbol{R\beta} = \boldsymbol{r}$$

There are three ways of doing that, but for now we will only discuss two. The first is to use the results in the previous section to infer that under the null,

$$(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r})'[\sigma_u^2 \boldsymbol{R}(\boldsymbol{X'X})^{-1}\boldsymbol{R}']^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r}) \sim \chi_q^2$$

where $\boldsymbol{R\beta}$ has been replaced by $\boldsymbol{r}$, the value of $\boldsymbol{R\beta}$ under the null. Again, if σ_u^2 is unknown, which is always the case i applications, we should instead use

$$F_1 = \frac{(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r})'[\boldsymbol{R}(\boldsymbol{X'X})^{-1}\boldsymbol{R}']^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r})/q}{\hat{\boldsymbol{u}}'\hat{\boldsymbol{u}}/(n-k)} \sim F_{q,n-k}.$$

which can also be written as

$$F_1 = (\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r})'[s_u^2 \boldsymbol{R}(\boldsymbol{X'X})^{-1}\boldsymbol{R}']^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r})/q \sim F_{q,n-k}.$$

We will call this is a *Wald-type test.*

The second way we can go about testing H_0 is to compare the residuals of the unrestricted model to the residuals of the model restricted so as the null is satisfied. Recall our discussion of Restricted LS, and let $\hat{u}$ be the residuals of the unrestricted model, and $\hat{u}_R$ be the residuals of the restricted model. Now observe that, under our assumptions, $\hat{u}'\hat{u} \sim \chi^2_{n-k}$, while $\hat{u}_R'\hat{u}_R \sim \chi^2_{n-k-q}$, so $(\hat{u}_R'\hat{u}_R - \hat{u}'\hat{u}) \sim \chi_q^2$ and

$$F_2 = \frac{(\hat{\boldsymbol{u}}_R'\hat{\boldsymbol{u}}_R - \hat{\boldsymbol{u}}'\hat{\boldsymbol{u}})/q}{\hat{\boldsymbol{u}}'\hat{\boldsymbol{u}}/(n-k)} \sim F_{q,n-k.}$$

We will call this an *LR-type test.*

Example 3. The hypothesis most frequently tested in applications is the "garbage regression" hypothesis

$$H_0 : \beta_2 = \beta_3 = \cdots = \beta_k = 0.$$

Under this hypothesis, all the *slope* coefficients in the regression are zero (the intercept β_1 is excluded from the list). If it is accepted, none of the regressors is important in explaining y, so the entire regression should be thrown into the garbage bin. The F statistic for this hypothesis is given by

$$F = \frac{n-k}{k-1} \cdot \frac{R^2}{1-R^2} \sim F_{k-1,n-k}.$$

The hypothesis is accepted when R^2, and thus F also, is close to 0, and rejected when R^2, and thus F also, is far from 0. ∎

12. Empirical Application: Estimating the Ellipticity of the Earth

Using the data of the *Academie des Sciences*, we estimate by least squares the regression model

$$\ell = \beta_0 + \beta_1(3\sin^2\theta) + u.$$

According to the Earth parameters in Table 2, the "true" equation is

$$\ell = 110.576 + 0.3723\ (3\sin^2\theta)$$

so that,

$$\begin{aligned} f &= 0.3723/110.576 = 1/297\ ; \\ C_E &= 110.576 \times 360 = 39,807\,\text{km}. \end{aligned}$$

The OLS estimates (along with their s.e.'s in parentheses below them) are

$$\begin{array}{ccccccccc} \hat{\ell} & = & 110.525 & + & 0.4697 & (3\sin^2\theta), & R^2 & = & 0.8773 \\ & & (0.158) & & (0.1014) & & \hat{\sigma}_u & = & 0.1903 \end{array}$$

so that,

$$\begin{aligned} \hat{f} &= 0.4697/110.525 = 1/235\ ; \\ \hat{C_E} &= 110.525 \times 360 = 39,789\,\text{km}. \end{aligned}$$

The variance-covariance matrix for $\hat{\beta}$ is given by

$$V(\hat{\beta}) = \hat{\sigma}_u^2(X'X)^{-1} = \begin{pmatrix} 0.02481 & -0.01344 \\ -0.01344 & 0.01028 \end{pmatrix}$$

Given the smallness of the sample size, the only way to perform inference is to assume that the neoclassical (normal and spherical errors) model applies. Figure 3 presents individual 95% CI's for β_0 and β_1, as well as, a joint 95% confidence region for the two parameters. This confidence region is obtained from equation (*) of Section 5.2 with $R = I$, $n = 5$, and $k = q = 2$, and is given by the set of values for which the quadratic form

$$\frac{1}{2}\begin{pmatrix} 110.525 - \beta_0 \\ 0.4697 - \beta_1 \end{pmatrix}' \begin{pmatrix} 0.02481 & -0.01344 \\ -0.01344 & 0.01028 \end{pmatrix}^{-1} \begin{pmatrix} 110.525 - \beta_0 \\ 0.4697 - \beta_1 \end{pmatrix}$$

is less than or equal to the critical $F_{1-\alpha,2,n-k} = F_{.95.2,3} = 9.552$ value.

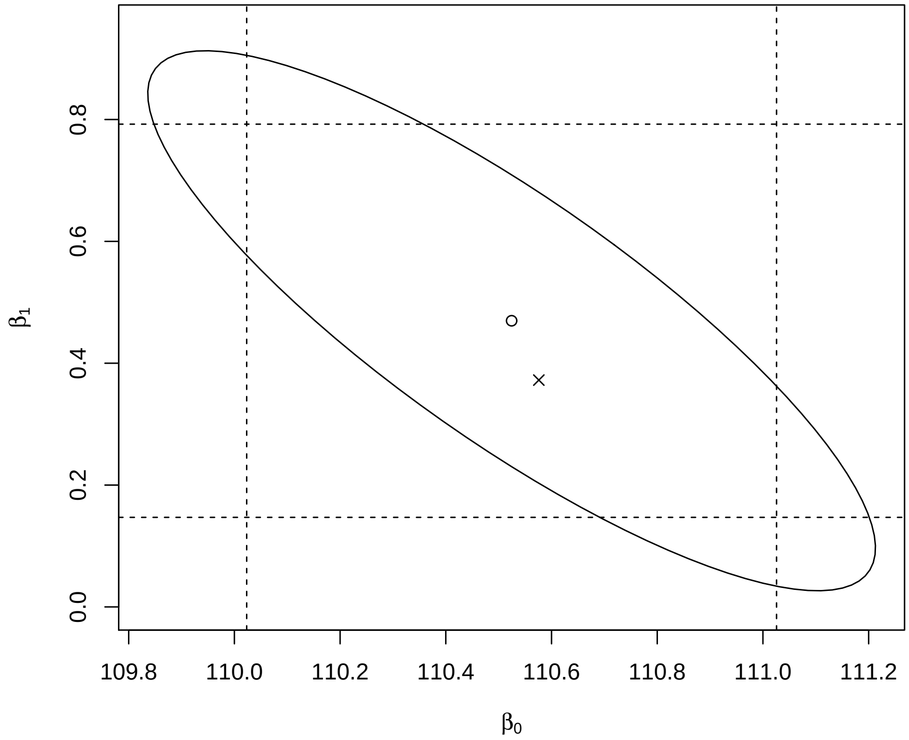

FIGURE 10. The 95% joint confidence region for β_0 and β_1 along with the marginal 95% CI's. The marginal 95% CI for β_0 is $[110.023, 111.026]$, while that for β_1 is $[0.1470, 0.7924]$. The now known true parameter values of $(\beta_0 = 110.576, \beta_1 = 0.3723)$ are also shown as point ×.

In order to obtain a confidence interval for $\psi = 1/f$ we resort to a trick: we divide both sides by $\hat{\beta}_0$ and estimate the OLS regression

$$\frac{\ell}{\hat{\beta}_0} = 1 + \frac{1}{\psi}(3\sin^2\theta) + \frac{u}{\hat{\beta}_0}.$$

Note that this regression has the same R^2 as the original regression and the t-statistic for $1/\psi$ is the same as the t-statistic for β_1. We get $1/\hat{\psi} = .0042498$ with an s.e. of .0009174, so, using the $t_{.975,3} = 3.182$ critical value, we obtain the 95% CI for $1/\psi$ as $[.001330, .007169]$. Upon inverting we get $\hat{\psi} = 235.3$ as above, and the 95% CI for ψ is given by

$$CI_{naive}(\psi; .95) = [139.48, 751.77].$$

This CI for ψ treats β_0 as known and equal to the estimated value without uncertainty, i.e., it does not take into account the variability in $\hat{\beta}_0$. Since, in our application, β_0 is estimated very accurately (i.e., it's s.e. is very small relative to its magnitude) this omission should not matter a lot. In any case, the correct 95% CI should be *wider* than this naive CI, so the latter can be thought of as a *lower bound* to the former.

To obtain the correct CI we use *Fieller's theorem*[7]. The following aside presents the method as it is adopted to the general linear model by Zerbe (1978)[8].

Aside. (Fieller's Theorem) Let

$$\psi = K\beta / L\beta,$$

where K and L are $1 \times k$ vectors of known constants, be the ratio of two linear combinations of a $k \times 1$ parameter vector β. If an estimator $\hat{\beta}$ is distributed as $\hat{\beta} \sim N(\beta, \Sigma)$, we have that, for a $\sqrt{n}$-estimator $\hat{\Sigma}$ of Σ,

$$T = \frac{K\hat{\beta} - \psi L\hat{\beta}}{\left[K\hat{\Sigma}K' - 2\psi K\hat{\Sigma}L' + \psi^2 L\hat{\Sigma}L'\right]^{1/2}} \sim t_{n-k}.$$

Letting $t = t_{1-\alpha/2,n-k}$ be the critical value from the t_{n-k} distribution, we have

$$1 - \alpha = \Pr\{-t \le T \le t\} = \Pr\{T^2 - t^2 \le 0\} = \Pr\{a\psi^2 + b\psi + c \le 0\},$$

where,

$$\begin{aligned} a &= (L\hat{\beta})^2 - t^2 L\hat{\Sigma}L', \\ b &= 2\left[t^2 K\hat{\Sigma}L' - (K\hat{\beta})(L\hat{\beta})\right], \\ c &= (K\hat{\beta})^2 - t^2 K\hat{\Sigma}K'. \end{aligned}$$

The last expression says that the interval containing the required $1-\alpha$ probability is characterized by the values for which the binomial $a\psi^2 + b\psi + c$ is *negative.* If a is positive, the function is convex and takes negative values. If, furthemore, the discriminant $b^2 - 4ac$ is positive, the binomial has 2 distinct real roots that define the required CI. Thus, the $100(1-\alpha)$% CI for ψ is given by

$$\left[\frac{-b - \sqrt{b^2 - 4ac}}{2a}, \frac{-b + \sqrt{b^2 - 4ac}}{2a}\right],$$

[7] Fieller, E.C. (1944), “A Fundamental Formula in the Statistics of Biology Assay and Some Applications”, *Quartely Journal of Pharmacy and Pharmacology*, 17, 117-123.

[8] Zerbe, G.O., (1978), “On Fieller's Theorem and the General Linear Model”, *The American Statistician*, 32, 103-105.

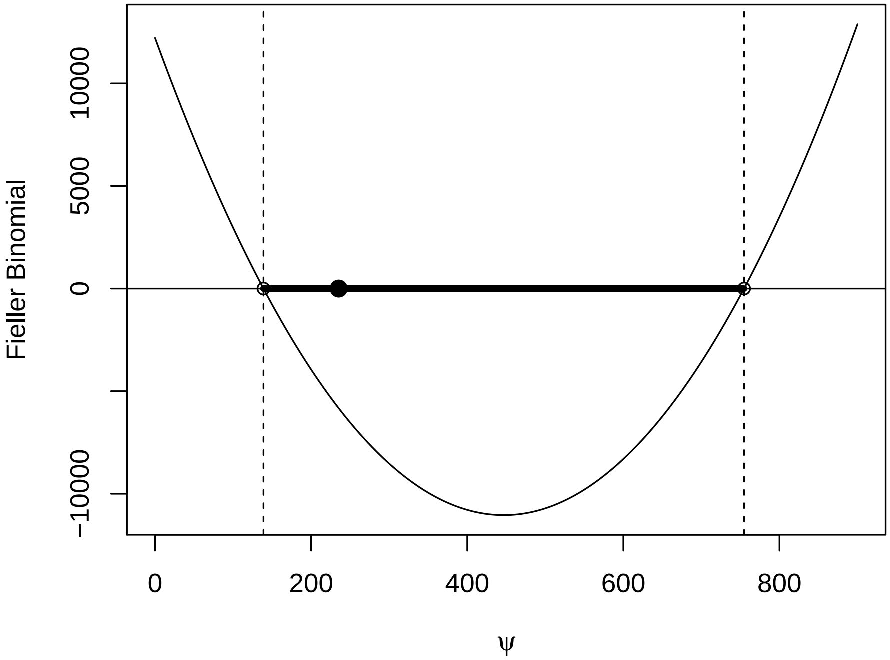

FIGURE 11. The Fieller binomial $a\psi^2 + b\psi + c$ (thin line) that, when negative, defines the Fieller 95% CI for ψ (bold line). The point estimate $\hat{\psi} = 235.3$ is also shown.

provided that $a > 0$ and $b^2 - 4ac > 0$. For the pathological $a < 0$ and/or $b^2 - 4ac < 0$ cases, as well as, for a nice geometrical interpretation of Fieller's theorem, see Luxburg and Franz (2009)[9]. ■

In our application, $\psi = \beta_0/\beta_1$, and $\hat{\beta}$ is the OLS coefficient that, under the normal and spherical errors assumption, is distributed as $N(\beta, \sigma_u^2(X'X)^{-1})$. Letting $K = (1, 0)$ and $L = (0, 1)$ we can write $\psi = K\beta/L\beta$ as required. Thus, using $t = t_{.975,3} = 3.182$, we compute

$$a = 0.1165 > 0, \quad b = -104.1, \quad c = 12,215.4$$
$$\text{and} \qquad b^2 - 4ac = 5,144.7 > 0.$$

[9] Luxburg, von U. and Franz, V.H., (2009), "A geometric approach to confidence sets for ratios: Fieller's theorem, generalizations and bootstrap", *Statistica Sinica*, 19, 1095-1117.

and the Fieller (correct) 95% CI for ψ is given by

$$CI_{Fieller}(\psi, .95) = [138.95, 754.66].$$

As expected, since β_0 is estimated quite accurately here, this correct CI for ψ (that takes into account the variability in both $\hat{\beta}_0$ and $\hat{\beta}_1$) is only marginally wider than the naive CI we computed above. Note that both CI's are assymetric around the point estimate $\hat{\psi} = 235.3$ (see Figure 9), with the longer tail towards large ψ's that correspond to a less oblate and more spherical earth.

Compared to the now known quantities, the estimates obtained by the French Academy of Sciences were indeed quite accurate. When published in the mid 1700's, these estimates lent considerable support to Newton's *Theory of Gravitation.* Ether, however, didn't go away, but was to play a crucial role in the development of the *Special Theory of Relativity* by Einstein, in the beginning of the 20th century. But that's another story.

13. The distribution of a truncated normal random variables

Consider two random variables X and Y that are jointly normally distributed with means μ_x and μ_y, variances σ_x^2 and σ_y^2, and correlation ρ. Their joint p.d.f. is given by

$$\begin{aligned} f^*(x,y) &= \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \times \\ &\quad \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_x}{\sigma_x}\right)^2 - 2\rho\left(\frac{x-\mu_x}{\sigma_x}\right)\left(\frac{y-\mu_y}{\sigma_y}\right) + \left(\frac{y-\mu_y}{\sigma_y}\right)^2\right]\right\}, \\ &\qquad -\infty < x < \infty, -\infty < y < \infty. \end{aligned}$$

Assume now that while we observe all values of X, we only observe Y if it is above a point a and below a point b, i.e., assume that Y is truncated from below at a and from above at b. Then the p.d.f. of X and the truncated Y is given by

$$f_{X,Y}(x,y) = \frac{f^*(x,y)}{\Phi\left(\frac{b-\mu_y}{\sigma_y}\right) - \Phi\left(\frac{a-\mu_y}{\sigma_y}\right)}, \qquad -\infty < x < \infty,\ a < y < b.$$

and 0 otherwise.

Direct integration gives the marginal density of X as

$$f_X(x) = \frac{\dfrac{1}{\sigma_x}\phi\left(\dfrac{x-\mu_x}{\sigma_x}\right)\left[\Phi\left(\dfrac{\dfrac{b-\mu_y}{\sigma_y} - \rho\dfrac{x-\mu_x}{\sigma_x}}{\sqrt{1-\rho^2}}\right) - \Phi\left(\dfrac{\dfrac{a-\mu_y}{\sigma_y} - \rho\dfrac{x-\mu_x}{\sigma_x}}{\sqrt{1-\rho^2}}\right)\right]}{\Phi\left(\dfrac{b-\mu_y}{\sigma_y}\right) - \Phi\left(\dfrac{a-\mu_y}{\sigma_y}\right)},$$

$$-\infty < x < \infty.$$

Defining $\alpha = (a-\mu_y)/\sigma_y$ and $\beta = (b-\mu_y)/\sigma_y$, we can write this density as

$$f_X(x) = \frac{1}{\sigma_x} g\left(\frac{x-\mu_x}{\sigma_x}\right), \qquad -\infty < x < \infty,$$

where

$$g(u) = \frac{\phi(u)\left[\Phi\left(\dfrac{\beta-\rho u}{\sqrt{1-\rho^2}}\right) - \Phi\left(\dfrac{\alpha-\rho u}{\sqrt{1-\rho^2}}\right)\right]}{\Phi(\beta)-\Phi(\alpha)}.$$

For $a = \mu_y$ and $b = \infty$, we have $\alpha = 0$ and $\beta = \infty$ and this becomes

$$g(u) = 2\phi(u)\Phi(\lambda u),$$

where $\lambda = \rho/\sqrt{1-\rho^2}$.

14. The distribution of the ratio of two normal variables

Let X and Y be two independent standard normal random variables and let

$$W = \frac{X+a}{Y+b}, \qquad a \geq 0, \quad b \geq 0 \tag{14.1}$$

be the ratio of the shifted random variables $X+a$ and $Y+b$, where a and b are non-negative constants.

We wish to express the distribution of W in terms of functions associated with measures of regions distribution function:

$$L(h,k,\rho) = \Pr[\xi > h, \eta > k]$$

where ξ and η are standard normal with covariance ρ, and the V function of Nicholson:

$$V(h,q) = \int_0^h \int_0^{qx/h} \phi(x)\phi(y)dydx,$$

where ϕ is the standard normal density. We have

$$\begin{aligned} P[W < t] &= P[X + a < t(Y + b), Y + b > 0] + P[X + a > t(Y + b), Y + b < 0] \\ &= P[-X + tY > a - bt, Y > -b] + P[X - tY > -a + bt, Y > b] \\ &= L\left(\frac{a - bt}{\sqrt{1+t^2}}\ , -b\ , \frac{t}{\sqrt{1+t^2}}\right) + L\left(\frac{-a + bt}{\sqrt{1+t^2}}\ , b\ , \frac{t}{\sqrt{1+t^2}}\right), \end{aligned}$$

since $Var(-X + tY) = V(X - tY) = 1 + t^2$, $Cov(-X + tY, Y) = Cov(X - tY, Y) = t$, and $Corr(-X + tY, Y) = Corr(X - tY, Y) = t/\sqrt{1+t^2}$. Then using the elementary properties of the L and V functions,

$$\begin{aligned} L(-h,-k,\rho) &= L(h,k,\rho) + \int_0^h \phi(x)dx + \int_0^k \phi(x)dx \\ L(-h,-k,\rho) + L(h,k,\rho) &= 2V\left(h, \frac{k - \rho h}{\sqrt{1-\rho^2}}\right) + 2V\left(k, \frac{h - \rho k}{\sqrt{1-\rho^2}}\right) \\ & \quad + \frac{1}{2} + \frac{\sin^{-1}\rho}{\pi}, \end{aligned}$$

we have several representations of $F(t) = P[W < t]$:

$$F(t) = L\left(\frac{a - bt}{\sqrt{1+t^2}}\ , -b\ , \frac{t}{\sqrt{1+t^2}}\right) + L\left(\frac{bt - a}{\sqrt{1+t^2}}\ , b\ , \frac{t}{\sqrt{1+t^2}}\right), \tag{14.2}$$

$$F(t) = \int_0^{(bt-a)/\sqrt{1+t^2}} \phi(x)dx + \int_0^b \phi(x)dx + 2L\left(\frac{bt - a}{\sqrt{1+t^2}}\ , b\ , \frac{t}{\sqrt{1+t^2}}\right), \tag{14.3}$$

$$F(t) = \frac{1}{2} + \frac{1}{\pi}\tan^{-1} t + 2V\left(\frac{bt - a}{\sqrt{1+t^2}}, \frac{b + at}{\sqrt{1+t^2}}\right) - 2V(b,a). \tag{14.4}$$

Representation (4) appears best for numerical purposes, unless b is large, say $b > 3$, since we have good methods for providing values of V. This last reference by D. B. Owen, also gives tables and formulas for the function

$$T(h,\lambda) = (2\pi)^{-1}\tan^{-1}\lambda - V(h, \lambda h),$$

which for some purposes is more convenient than the V function.

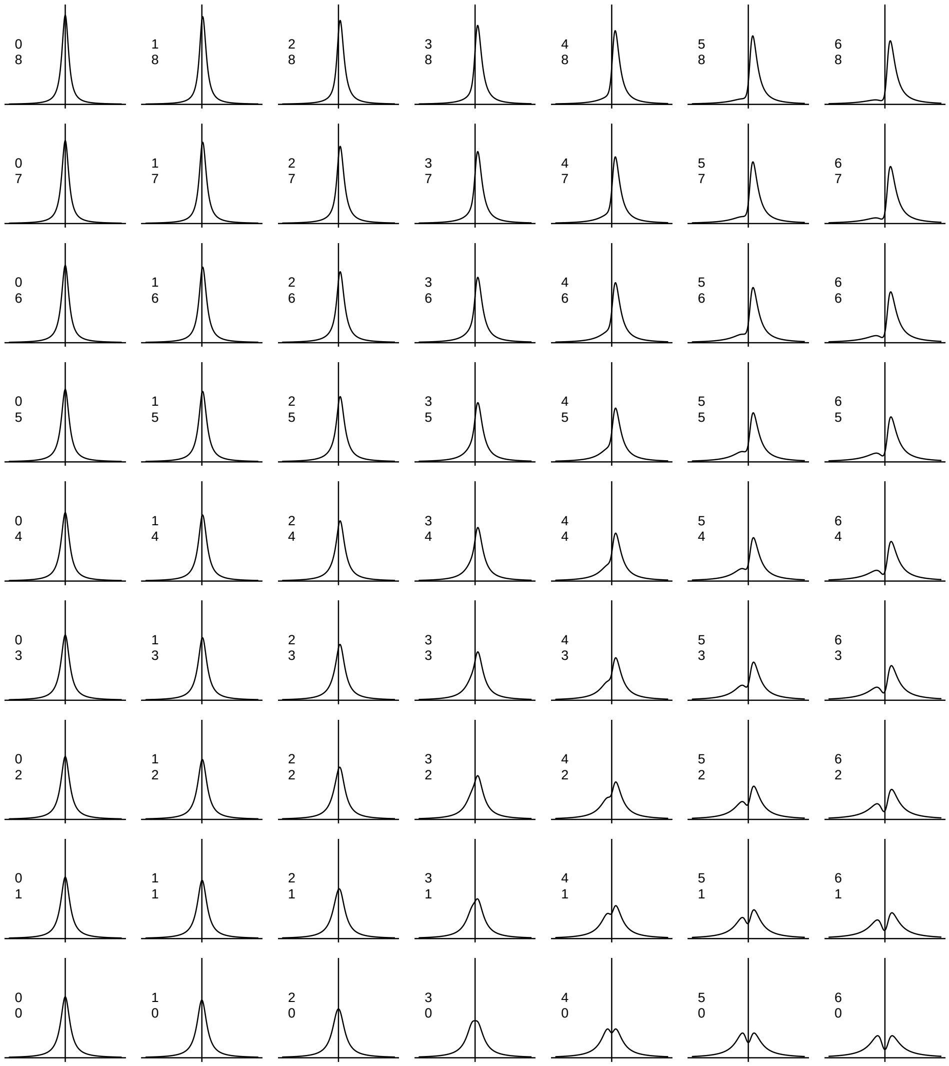

FIGURE 12. The density of $(X+a)/(Y+b)$ for various values of a and b. Here $a =$ upper number/3 and $b =$ lower number/8.

15. MODEL SELECTION

The F-statistic for testing the joint significance of the complete set of regressors is

$$F = \frac{\text{ESS}/(k-1)}{\text{RSE}/(n-k)} = \frac{R^2/(k-1)}{(1-R^2)/(n-k)} \sim F_{(k-1),n-k}.$$

There are 3 popular model selection criteria:

1. Maximize *Adjusted* R^2

$$\bar{R}^2(k) = 1 - \frac{\text{RSS}/(n-k)}{\text{TSS}/(n-1)}.$$

2. Minimize *Akaike Criterion*

$$\text{AIC}(k) = \log \frac{\hat{u}'\hat{u}}{n} + \frac{2k}{n}.$$

3. Minimize *Schwarz criterion*

$$\text{BIC}(k) = \log \frac{\hat{u}'\hat{u}}{n} + \frac{k}{n} \log n.$$

> THE
> NORMAL
> LAW OF ERROR
> STANDS OUT IN THE
> EXPERIENCE OF MANKIND
> AS ONE OF THE BROADEST
> GENERALIZATIONS OF NATURAL
> PHILOSOPHY ♦ IT SERVES AS THE
> GUIDING INSTRUMENT IN RESEARCHES
> IN THE PHYSICAL AND SOCIAL SCIENCES AND
> IN MEDICINE AGRICULTURE AND ENGINEERING ♦
> IT IS AN INDISPENSABLE TOOL FOR THE ANALYSIS AND THE
> INTERPRETATION THE BASIC DATA OBTAINED BY OBSERVATION AND EXPERIMENT
>
> — *W. J. Youden*

> Everyone believes in the Gaussian law of errors, the experimenters because they think it is a mathematical theorem, the mathematicians because they think it is an experimental fact.
>
> — *Henri Poincaré*