A2. Statistics from a Geometric Viewpoint

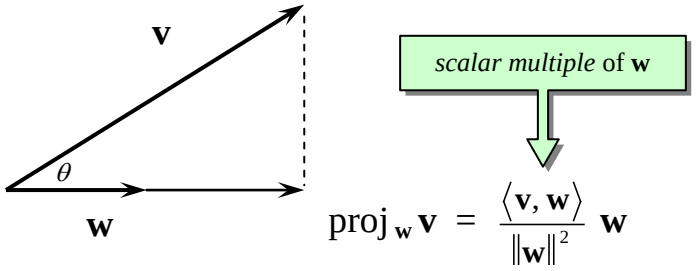
Least Squares Approximation

The concepts of linear correlation and least squares regression can be viewed very elegantly, from a pure geometric perspective. Again, recall some basic background facts from elementary vector analysis:

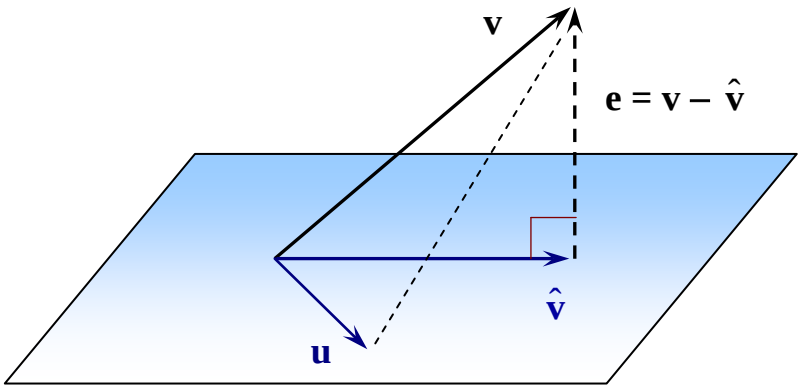
For any two column vectors $\mathbf{v} = (v_1, v_2, \ldots, v_n)^T$ and $\mathbf{w} = (w_1, w_2, \ldots, w_n)^T$ in $\mathbb{R}^n$, the standard Euclidean *dot product* "$\mathbf{v} \cdot \mathbf{w}$" is defined as $\mathbf{v}^T \mathbf{w} = \sum_{i=1}^{n} v_i w_i$, hence is a scalar. Technically, the dot product is a special case of a more general mathematical object known as an *inner product*, denoted by $\langle \mathbf{v}, \mathbf{w} \rangle$, and these notations are often used interchangeably. The length, or *norm*, of a vector $\mathbf{v}$ can therefore be characterized as $\|\mathbf{v}\| = \sqrt{\langle \mathbf{v}, \mathbf{v} \rangle} = \sqrt{\sum_{i=1}^{n} v_i^2}$, and the included angle θ between two vectors $\mathbf{v}$ and $\mathbf{w}$ can be calculated via the formula

$$\cos\theta = \frac{\langle \mathbf{v}, \mathbf{w} \rangle}{\|\mathbf{v}\| \, \|\mathbf{w}\|}, \qquad 0 \le \theta \le \pi.$$

From this relation, it is easily seen that two vectors $\mathbf{v}$ and $\mathbf{w}$ are *orthogonal* (i.e., $\theta = \pi/2$), written $\mathbf{v} \perp \mathbf{w}$, if and only if their dot product is equal to zero, i.e., $\langle \mathbf{v}, \mathbf{w} \rangle = 0$. More generally, the *orthogonal projection* of the vector $\mathbf{v}$ onto the vector $\mathbf{w}$ is given by the formula shown in the figure below. (Think of it informally as the "shadow vector" that $\mathbf{v}$ casts in the direction of $\mathbf{w}$.)



Why are orthogonal projections so important? Suppose we are given any vector $\mathbf{v}$ (in a general *inner product space*), and a plane (or more precisely, a *linear subspace*) not containing $\mathbf{v}$. Of all the vectors $\mathbf{u}$ in this plane, we wish to find a vector $\hat{\mathbf{v}}$ that comes "closest" to $\mathbf{v}$, in some formal mathematical sense. The **Best Approximation Theorem** asserts that, under such very general conditions, such a vector does indeed exist, and is uniquely determined by the orthogonal projection of $\mathbf{v}$ onto this plane. Moreover, the resulting error $\mathbf{e} = \mathbf{v} - \hat{\mathbf{v}}$ is smallest possible, with $\|\mathbf{e}\|^2 = \|\mathbf{v}\|^2 - \|\hat{\mathbf{v}}\|^2$, via the Pythagorean Theorem.



Of all the vectors $\mathbf{u}$ in the plane, the one that minimizes the length $\|\mathbf{v} - \mathbf{u}\|$ (thin dashed line) is the orthogonal projection $\hat{\mathbf{v}}$. Therefore, $\hat{\mathbf{v}}$ is the *least squares approximation* to $\mathbf{v}$, yielding the *least squares error* $\|\mathbf{e}\|^2 = \|\mathbf{v}\|^2 - \|\hat{\mathbf{v}}\|^2$.

Now suppose we are given n data points (x_i, y_i), $i = 1, 2, \ldots, n$, obtained from two variables X and Y. Define the following vectors in n-dimensional Euclidean space $\mathbb{R}^n$:

$\mathbf{0} = (0, 0, 0, \ldots, 0)^T$, $\quad \mathbf{1} = (1, 1, 1, \ldots, 1)^T$,

$\mathbf{x} = (x_1, x_2, x_3, \ldots, x_n)^T$, $\quad \bar{\mathbf{x}} = (\bar{x}, \bar{x}, \bar{x}, \ldots, \bar{x})^T$, so that $\mathbf{x} - \bar{\mathbf{x}} = (x_1 - \bar{x}, x_2 - \bar{x}, x_3 - \bar{x}, \ldots, x_n - \bar{x})^T$,

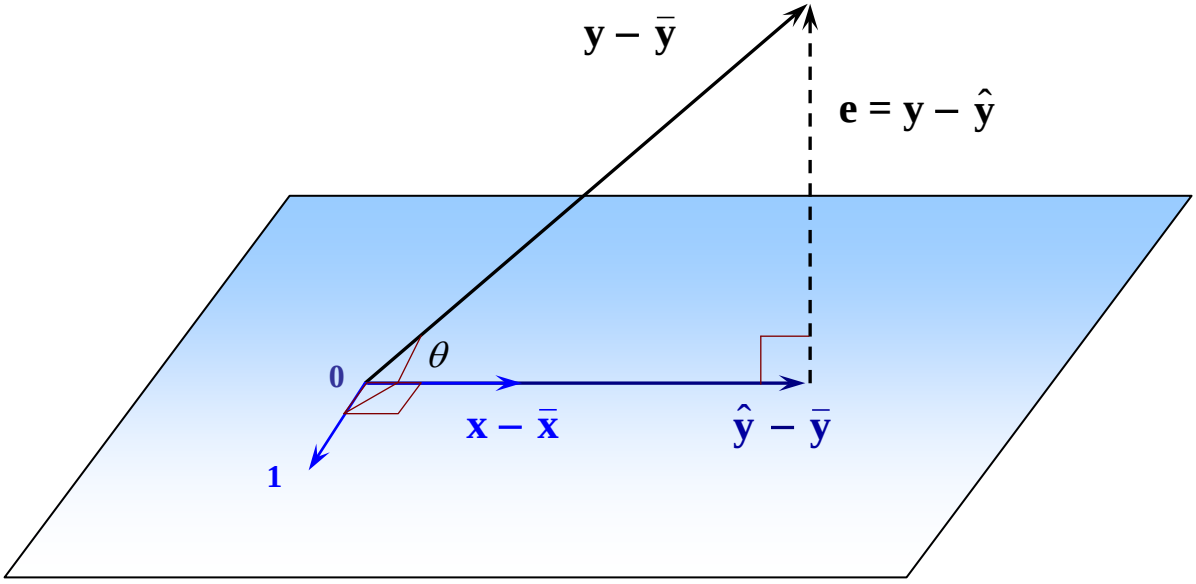
$\mathbf{y} = (y_1, y_2, y_3, \ldots, y_n)^T$, $\quad \bar{\mathbf{y}} = (\bar{y}, \bar{y}, \bar{y}, \ldots, \bar{y})^T$, so that $\mathbf{y} - \bar{\mathbf{y}} = (y_1 - \bar{y}, y_2 - \bar{y}, y_3 - \bar{y}, \ldots, y_n - \bar{y})^T$.

The "centered" data vectors $\mathbf{x} - \bar{\mathbf{x}}$ and $\mathbf{y} - \bar{\mathbf{y}}$ are crucial to our analysis. For observe that, by definition,

$$\|\mathbf{x} - \bar{\mathbf{x}}\|^2 = (n-1)\, s_x^2, \qquad \|\mathbf{y} - \bar{\mathbf{y}}\|^2 = (n-1)\, s_y^2, \qquad \text{and} \qquad \langle \mathbf{x} - \bar{\mathbf{x}}, \mathbf{y} - \bar{\mathbf{y}} \rangle = (n-1)\, s_{xy}.$$

Now, note that $\langle \mathbf{1}, \mathbf{x} - \bar{\mathbf{x}} \rangle = \sum_{i=1}^{n} (x_i - \bar{x}) = 0$, therefore $\mathbf{1} \perp (\mathbf{x} - \bar{\mathbf{x}})$; likewise, $\mathbf{1} \perp (\mathbf{y} - \bar{\mathbf{y}})$ as well.

See the figure below, showing the geometric relationships between the vector $\mathbf{y} - \bar{\mathbf{y}}$ and the plane *spanned by the orthogonal basis vectors* $\mathbf{1}$ and $\mathbf{x} - \bar{\mathbf{x}}$.



Also, from a previous formula, we see that the general angle θ between these two vectors is given by

$$\cos\theta = \frac{\langle \mathbf{x} - \bar{\mathbf{x}}, \mathbf{y} - \bar{\mathbf{y}} \rangle}{\|\mathbf{x} - \bar{\mathbf{x}}\| \, \|\mathbf{y} - \bar{\mathbf{y}}\|}$$

(from above)

$$= \frac{(n-1)\, s_{xy}}{\sqrt{(n-1)\, s_x^2}\, \sqrt{(n-1)\, s_y^2}} = \boxed{\frac{s_{xy}}{s_x\, s_y} = r}$$

i.e., the sample linear correlation coefficient! Therefore, this ratio r measures the cosine of the angle θ between the vectors $\mathbf{x} - \bar{\mathbf{x}}$ and $\mathbf{y} - \bar{\mathbf{y}}$, and hence is always between –1 and +1. But what is its exact connection with the original vectors $\mathbf{x}$ and $\mathbf{y}$?

IF the vectors $\mathbf{x}$ and $\mathbf{y}$ are exactly *linearly correlated*, then by definition, it must hold that $\boxed{\mathbf{y} = b_0\,\mathbf{1} + b_1\,\mathbf{x}}$ for some constants b_0 and b_1, and conversely. A little elementary algebra (take the mean of both sides, then subtract the two equations from one another) shows that this is equivalent to the statement

$$\mathbf{y} - \bar{\mathbf{y}} = b_1\,(\mathbf{x} - \bar{\mathbf{x}}), \qquad \text{with} \quad \boxed{b_0 = \bar{y} - b_1\,\bar{x}}.$$

That is, the vector $\mathbf{y} - \bar{\mathbf{y}}$ is a *scalar multiple* of the vector $\mathbf{x} - \bar{\mathbf{x}}$, and therefore must lie not only in the plane, but along the line *spanned* by $\mathbf{x} - \bar{\mathbf{x}}$ itself. If the scalar multiple $b_1 > 0$, then $\mathbf{y} - \bar{\mathbf{y}}$ must point in the same direction as $\mathbf{x} - \bar{\mathbf{x}}$; hence $r = \cos 0 = +1$, and the linear correlation is positive. If $b_1 < 0$, then these two vectors point in opposite directions, hence $r = \cos \pi = -1$, and the linear correlation is negative. However, if these two vectors are orthogonal, then $r = \cos(\pi/2) = 0$, and there is no linear correlation between $\mathbf{x}$ and $\mathbf{y}$.

More generally, if the original vectors $\mathbf{x}$ and $\mathbf{y}$ are not exactly linearly correlated (that is, $-1 < r < +1$), then the vector $\mathbf{y} - \bar{\mathbf{y}}$ does not lie in the plane. The unique vector $\hat{\mathbf{y}} - \bar{\mathbf{y}}$ that *does* lie in the plane which best approximates it in the "least squares" sense is its orthogonal projection onto the vector $\mathbf{x} - \bar{\mathbf{x}}$, computed by the formula given above:

$$\hat{\mathbf{y}} - \bar{\mathbf{y}} = \frac{\langle \mathbf{y} - \bar{\mathbf{y}},\, \mathbf{x} - \bar{\mathbf{x}} \rangle}{\|\mathbf{x} - \bar{\mathbf{x}}\|^2}\,(\mathbf{x} - \bar{\mathbf{x}})$$

$$= \frac{(n-1)\,s_{xy}}{(n-1)\,s_x^{\,2}}\,(\mathbf{x} - \bar{\mathbf{x}}),$$

i.e., **Linear Model:** $\boxed{\hat{\mathbf{y}} - \bar{\mathbf{y}} = b_1\,(\mathbf{x} - \bar{\mathbf{x}})}$, with $\boxed{b_1 = \dfrac{s_{xy}}{s_x^{\,2}}}$.

Furthermore, via the Pythagorean Theorem,

$$\|\mathbf{y} - \bar{\mathbf{y}}\|^2 = \|\hat{\mathbf{y}} - \bar{\mathbf{y}}\|^2 + \|\mathbf{y} - \hat{\mathbf{y}}\|^2$$

or, in statistical notation… $\boxed{SS_{Total} = SS_{Reg} + SS_{Error}}$.

Finally, from this, we also see that the ratio

$$\frac{SS_{Reg}}{SS_{Total}} = \frac{\|\hat{\mathbf{y}} - \bar{\mathbf{y}}\|^2}{\|\mathbf{y} - \bar{\mathbf{y}}\|^2}$$

$$= \cos^2\theta,$$

i.e., the **coefficient of determination** is

$\boxed{\dfrac{SS_{Reg}}{SS_{Total}} = r^2}$, where r is the **correlation coefficient**.

Exercise: Derive the previous formulas $s_{x \pm y}^{\,2} = s_x^{\,2} + s_y^{\,2} \pm 2 s_{xy}$. (*Hint*: Use the Law of Cosines.)

Remark: In this analysis, we have seen how the familiar formulas of linear regression follow easily and immediately from "orthogonal approximation" on vectors. With slightly more generality, interpreting vectors abstractly as functions $f(x)$, it is possible to develop the formulas that are used in **Fourier series**.