CORRELATION AND NORMAL DISTRIBUTIONS

p. 134, another formula for r, the correlation: If the data points are (x_i, y_i) for $i = 1, \dots, n$, then we defined r on p. 132 as

$$r := \frac{1}{n} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{\sigma_x} \right) \left(\frac{y_i - \bar{y}}{\sigma_y} \right),$$

where $\bar{x}$ is the mean of x_i and σ_x is the sd of x_i. (We'll use "sd" for the standard deviation of data and "SD" for the standard deviation of random variables. Likewise for "var"/"Var" and "cov"/"Cov".) We can use algebra to rewrite this as the formula on p. 134:

$$\begin{aligned} r &= \frac{1}{\sigma_x \sigma_y} \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{\sigma_x \sigma_y} \frac{1}{n} \left[\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \bar{y} - \sum_{i=1}^{n} \bar{x} y_i + \sum_{i=1}^{n} \bar{x} \bar{y} \right] \\ &= \frac{1}{\sigma_x \sigma_y} \frac{1}{n} \left[\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} - n\bar{x}\bar{y} + n\bar{x}\bar{y} \right] = \frac{1}{\sigma_x \sigma_y} \left[\frac{1}{n} \sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y} \right] = \frac{\operatorname{cov}(x, y)}{\sigma_x \sigma_y}, \end{aligned}$$

as claimed.

p. 147, typical vertical distance to the SD line: The SD line goes through $(\bar{x}, \bar{y})$ and has slope σ_y/σ_x if $r \geq 0$, or slope $-\sigma_y/\sigma_x$ if $r \leq 0$. Therefore, the equation of the SD line is

$$y - \bar{y} = \frac{\sigma_y}{\sigma_x}(x - \bar{x})$$

if $r \geq 0$ (if $r < 0$, then the right-hand side is multiplied by -1). We could also write this as $y = \bar{y} + (\sigma_y/\sigma_x)(x - \bar{x})$. The point on the SD line corresponding to x_i therefore has y-coordinate $\bar{y} + (\sigma_y/\sigma_x)(x_i - \bar{x})$, so the vertical distance to (x_i, y_i), being the difference of the y coordinates, is $y_i - \big[\bar{y} + (\sigma_y/\sigma_x)(x_i - \bar{x})\big]$. Thus, the *square of* the r.m.s. vertical distance is

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^{n} \Big(y_i - [\bar{y} + (\sigma_y/\sigma_x)(x_i - \bar{x})] \Big)^2 &= \frac{1}{n} \sum_{i=1}^{n} \Big([y_i - \bar{y}] - [(\sigma_y/\sigma_x)(x_i - \bar{x})] \Big)^2 \\ &= \frac{1}{n} \sum_{i=1}^{n} \Big([y_i - \bar{y}]^2 + [(\sigma_y/\sigma_x)(x_i - \bar{x})]^2 - 2[y_i - \bar{y}][(\sigma_y/\sigma_x)(x_i - \bar{x})] \Big) \\ &= \sigma_y^2 + (\sigma_y/\sigma_x)^2 \sigma_x^2 - \frac{2(\sigma_y/\sigma_x)}{n} \sum_{i=1}^{n} [y_i - \bar{y}][(x_i - \bar{x})] \\ &= 2\sigma_y^2 - 2\sigma_y^2 r = 2(1 - r)\sigma_y^2, \end{aligned}$$

as claimed. A similar calculation works for $r < 0$, giving $-r$ in place of r. Thus, both expressions are equal to $2\big(1 - |r|\big)\sigma_y^2$.

p. 186, formula for r.m.s. regression error and p. 208, line of best fit: We'll standardize, so $\bar{x} = \bar{y} = 0$ and $\sigma_x = \sigma_y = 1$. We want to know which line $y = mx + b$ minimizes

$(1/n)\sum_{i=1}^n [y_i-(mx_i+b)]^2$. If we expand this sum, we get $1+m^2+b^2-2mr = 1-r^2+b^2+(m-r)^2$. Now it is clear that the minimum occurs at $b=0$ and $m=r$. Furthermore, the squared r.m.s. error is then $1-r^2$, so the r.m.s. error is then $\sqrt{1-r^2}$. If we don't standardize, this becomes $\sqrt{1-r^2}\sigma_y$.

The way to analyze jointly normal distributions is usually not by using the density (in fact, some don't even have densities), but by using the following definition and key theorem: A sequence $X_1,\dots,X_n$ is called **jointly normal** or **multivariate normal** if there exist independent normal random variables $Y_1,\dots,Y_m$ so that each X_i is a linear combination of the Y_j: $X_i = \sum_{j=1}^m a_{i,j}Y_j$ for some constants $a_{i,j}$. (This can be expressed with matrices nicely, which we'll use later. Namely, if X is a column vector with entries X_i, Y is a column vector with entries Y_j, and A is a matrix with entries $a_{i,j}$, then $X=AY$. It follows that for every matrix B with n columns, BX is normal since $BX = (BA)Y$.) Note that a constant random variable is normal; it just has variance 0. A key theorem about multivariate normal distributions is that they are uniquely determined by their means $E(X_i)$ and covariances $E(X_iX_k)-E(X_i)E(X_k)$. In other words, if two jointly normal sequences have the same means and covariances, then they have the same distributions. The most well-known special case of this is that if the covariances between pairs of different random variables are all 0, then the random variables are independent. (This is certainly not true of other random variables!)

For example, suppose we want a bivariate normal (X,Y) with means μ_X and μ_Y, SDs σ_X and σ_Y, and with correlation r. Define $s := \sqrt{1-r^2}$. Let Z_1, Z_2 be independent standard normal random variables (that is, they have mean 0 and variance 1). Define $X := \mu_X + \sigma_X Z_1$ and $Y := \mu_Y + \sigma_Y(rZ_1 + sZ_2)$. You should be able to check that (X,Y) has the desired properties by using the fact that $r^2+s^2=1$. Note that many aspects are easier to see when we standardize, which will make $X=Z_1$ and $Y = rZ_1+sZ_2$.

---

p. 160, the regression method: As long as the data are normally distributed, this now follows easily by the following calculation:

$$\frac{E(Y \mid X)-\mu_Y}{\sigma_Y} = \frac{E\big(\sigma_Y(rZ_1+sZ_2) \mid Z_1\big)}{\sigma_Y} = rZ_1 = r\frac{X-\mu_X}{\sigma_X}\,.$$

If we had standardized, this would have been much shorter: $E(Y \mid X) = rX$.

p. 182, rule of thumb for r.m.s. regression error: As long as the data are normally distributed, this now follows easily by the following calculation: The regression line has the equation $y-\mu_Y = r(\sigma_Y/\sigma_X)(x-\mu_X)$, so the distribution of the vertical distance to the regression line is the distribution of $Y-[\mu_Y + r(\sigma_Y/\sigma_X)(X-\mu_X)] = Y-\mu_Y-\sigma_Y rZ_1 = \sigma_Y sZ_2$, i.e., $\mathrm{N}(0,\sigma_Y^2 s^2)$. In particular, the r.m.s. error is $\sigma_Y s$ and is the same even when given X (homoscedasticity) since Z_2 is independent of Z_1 and hence of X. Since the distance is normally distributed, the rule of thumb follows. Again, if we had standardized, this would have been much easier: the regression line is then $y=rx$ and the distance has the distribution of $Y-rX = sZ_2$.