Family Regression Examples

Dr. Scott

August 29, 2005

The multivariate normal distribution has a number of closed-form expressions of interest. We'll not pursue them all here, but for your interest, here is one of the more powerful results (cf, Stat 541, Multivariate Stats).

In class, we will generally assume that X is chosen, then Y measured. Here, we assume both X and Y are random from a multivariate normal distribution, $N(\mu, \Sigma)$, where $z = (x, y)$:

$$f(z) = |2\pi\Sigma|^{-1/2} \exp[-\frac{1}{2}(z-\mu)^t \Sigma^{-1}(z-\mu)].$$

More explicitly, $f(z) = f(x, y) \sim N(\mu, \Sigma)$ can be decomposed as follows:

$$\mu = \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix} \qquad \text{and} \qquad \Sigma = \begin{pmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{pmatrix}.$$

Here the marginal densities are also normal:

$$\begin{aligned} f(x) &\sim N(\mu_x, \Sigma_{xx}) \qquad \text{and} \\ f(y) &\sim N(\mu_y, \Sigma_{yy}). \end{aligned}$$

Now it is always the case that

$$f(z) = f(x, y) = f(x) f(y|x).$$

What is $f(y|x)$? It is an amazing algebraic fact that

$$f(y|x) \sim N(\mu_{y|x}, \Sigma_{yy.x})$$

where

$$\begin{aligned} \mu_{y|x} &= \mu_y - \Sigma_{yx}\Sigma_{xx}^{-1}\mu_x \\ \Sigma_{yy.x} &= \Sigma_{yy} - \Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}. \end{aligned}$$

The simple linear regression case corresponds to both x and y one-dimensional, where

$$\Sigma = \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix}$$

so that

$$\begin{aligned} \Sigma_{yy.x} &= \Sigma_{yy} - \Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy} \\ &= \sigma_y^2 - (\rho\sigma_x\sigma_y)\cdot(\sigma_x^2)^{-1}\cdot(\rho\sigma_x\sigma_y) \\ &= \sigma_y^2 - \rho^2\sigma_y^2 \\ &= (1-\rho^2)\sigma_y^2 \\ &\equiv \sigma_{y|x}^2. \end{aligned}$$

As Pearson discovered, the correlation between any pair of parent/child heights is exactly $\rho = 1/2$. The same is true of siblings. The correlation between mother and father is approximately $\rho = 1/4$.

The standard deviation of height is about $2.5'' = \sigma_y$. If you know a parent's height, then $1 - \rho^2 = 3/4$ and

$$\sigma_{y|x} = \sigma_y \cdot \sqrt{(1 - \rho^2)} = \sigma_y \cdot 0.866,$$

which is a 14% reduction in uncertainty.

How much can you improve your prediction if you know both parents' heights? Well,

$$\Sigma = \begin{pmatrix} 1 & \frac{1}{4} & \frac{1}{2} \\ \frac{1}{4} & 1 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 1 \end{pmatrix}$$

Using the formula for $\Sigma_{yy.x}$,

$$\sigma^2_{y|x} = \sigma^2_y \frac{3}{5} \qquad \text{or} \qquad \sigma_{y|x} = \sigma_y \cdot 0.77$$

a 23% reduction in uncertainty.

If you add information about the heights of your siblings to your parents, that will also improve your prediction accuracy. That includes your younger siblings. In standard units, a variance of 1 becomes

0.75 with one parent

0.6 with two parents

0.5833 with two parents and one sibling

0.5713 with two parents and two siblings

0.5625 with two parents and three siblings

You can also use the children's heights to predict your parents' heights! Thus you can choose to predict things that are clearly not causal.

It is an easily forgotten lesson, but correlation does not prove causation. (However, lack of correlation does support the lack of causation, at least in a linear sense.)