Bivariate Normal Distribution, the Regression Effect

A link between correlation and regression: \( r^2 \) measures fraction of variation explained

The “regression effect”:
Galton’s fathers and sons.
A chance error model for test-retest

Bivariate Normal Distribution: Parametric interpretation, the linear model.

- A joint distribution for two random variables \((X,Y)\)
- Specified by five population parameters:
  - Means \( \left( \begin{array}{c} \mu_X \\ \mu_Y \end{array} \right) \)
  - Variances \( \left( \begin{array}{c} \sigma_X^2 \\ \sigma_Y^2 \end{array} \right) \)
  - Correlation \( \rho \)

- The joint probability density function is complicated, and we won’t need it, except to point out that it is a bivariate bell shape. It is a nice rotationally symmetric bell if \( \rho = 0 \), but otherwise it is elongated/squashed by an amount depending on \( -1 \leq \rho \leq 1 \).

- The marginal distributions are normal: \( X \sim N(\mu_X, \sigma_X^2) \), \( Y \sim N(\mu_Y, \sigma_Y^2) \)

But these marginal distributions don’t reveal the role of the correlation \( \rho \): to see this, we need to look at:

- The conditional distributions are normal also:
  \[
  \mu_{Y|x} = E[Y|X = x] = \mu_Y + \rho \frac{\sigma_Y}{\sigma_X} (x - \mu_X)
  \]
  \[
  \sigma_{Y|x}^2 = Var(Y|X = x) = (1 - \rho^2)\sigma_Y^2
  \]

The first of these equations is the population regression line of \( Y \) on \( X \). It is a population version of the graph of averages.

Example (from book, page 515):

```r
> rata <- matrix(c(112.6, 90.0, 90.0, 102.1, 90.0, 77.1, 77.1, 81.5, 81.5, 57.1, 75.5, 73.3, 75.5, 73.3, 67.3, 67.3, 55.3, 55.3, 57.1, 57.1, 55.3, 57.1), nrow=8, byrow=TRUE)
> tapply(rata[,1],rata[,2],sort)
      0       2.5       5
81.5 90.0 93.0 102.1 105.6 106.6 108.3 112.6
> tapply(rata[,1],rata[,2],mean)
      0       2.5       5
99.9875 75.5000 54.9500
> tapply(rata[,1],rata[,2],var)
      0       2.5       5
113.467 115.209 176.923
plot(rata[,2],rata[,1])
```
\[ \mu_{Y|X=0} = 100, \quad \mu_{Y|X=2.5} = 75.5, \quad \mu_{Y|X=2.5} = 55 \]

\[ \sigma_{Y|X=0}^2 = 113, \quad \sigma_{Y|X=2.5}^2 = 115, \quad \sigma_{Y|X=2.5}^2 = 177 \]

(Note also, the same is true interchanging the roles of \(X\) and \(Y\), the distribution of \(X\) given \(Y=y\) is also normal with mean and variance given by)

\[ \mu_{X|y} = E[X|Y = y] = \mu_x + \rho \frac{\sigma_x}{\sigma_y} (y - \mu_y) \]

\[ \sigma_{X|y}^2 = (1 - \rho^2) \sigma_x^2 \]

**Connection with least squares:** We may equivalently write the regression line in the form

\[ \hat{Y}(x) = \alpha + \beta x \]

where the (population) slope and intercept coefficients are given by

\[ \alpha = \mu_y - \beta \mu_x, \quad \beta = \rho \frac{\sigma_y}{\sigma_x} \quad (+) \]

These are just the population versions of the least squares coefficients from a sample:

\[ a = \bar{y} - b \bar{x}, \quad b = r \frac{s_y}{s_x} \]

rat.lm=lm( rata[, 1] ~ rata[, 2])

lm(formula = rata[, 1] ~ rata[, 2])

Coefficients:

(Intercept) rata[, 2]
    99.331        -9.008

fitted(rat.lm)

1 2 3 4 5 6 7 8
9 10 11 12 13 14 15 16
76.81250 76.81250 76.81250 76.81250 76.81250 76.81250 76.81250 76.81250
2. CORRELATION VS. REGRESSION

In correlation, x and y play similar roles. In regression, one (x) is explanatory and the other (y) is a response. Indeed, in regression, if we reversed the role of explanatory and response, we would get a different regression line.

Another link between correlation and regression: \( r^2 = \text{fraction of variation in Y explained by the regression} \)

2 sources of variation in y (weight):

1. x (height) → \( \hat{y} = a + bx \) → variation in predicted \( \hat{y} \).
2. given x (fixed vertical strip), variation in weight given height (y|x) \( r^2 = \frac{\text{Variance of predicted } \hat{y}}{\text{Variance of predicted } y} = \frac{s_{\hat{y}}^2}{s_y^2} \)

[ \( r^2 \) can be pretty small! In heights and weights slide, \( r^2 = (.47)^2 = 0.22 \)]

Population version of this discussion. (Same ideas, but now expressed in terms of the population, rather than the sample). We again decompose the variation of Y into two components: variation due to X, and variation remaining in Y after fixing X.

a) Variation due to X. Recall that

\[ \hat{Y}(X) = \mu_{Y|X} = \alpha + \beta X, \text{ with } \beta = \rho \frac{\sigma_Y}{\sigma_X}. \]

From the formula for variance of a linear combination

\[ \text{Var } \hat{Y}(X) = \beta^2 \text{Var } X = \rho^2 \frac{\sigma_Y^2}{\sigma_X^2} \times \sigma_X^2 = \rho^2 \sigma_Y^2. \]

b) Variance left in Y given X:

\[ \text{Var } (Y|X = x) = (1 - \rho^2)\sigma_Y^2. \]

Adding together the variation from the two sources a) and b), we have

\[ \text{Var } Y = \sigma_Y^2 = \rho^2 \sigma_Y^2 + (1 - \rho^2)\sigma_Y^2 = \text{Var } \hat{Y}(X) + \text{Var } (Y|X) \]

This leads to the interpretation

\[ \rho^2 = \frac{\text{Var } \hat{Y}(X)}{\text{Var } Y} = \text{fraction of variation in “explained” by the regression} \]

3. THE REGRESSION EFFECT

Preliminary question: Let be i’th student’s scores on quizzes 1 and 2. Suppose X and Y have mean 100, SD 15, and correlation \( r = .6 \)
A claim: (i) “The scores of people above average on the first test will drop overall by 5 on the second test”
(ii) “The scores of people below average on the first test will rise overall by 5 on the second test”
Is this right? Is there a reason? Or just chance? [Substitutes: IQ tests, mutual fund returns . . . .]

Why “regression”? Go back to Galton and father-son pairs.

\[
\bar{x} = 68 \quad s_x = 2.7 \quad r = 0.5
\]

Consider 72” fathers +4” above average. You might think that their sons would be 4” above average too, i.e. 73” – wrong! Since most are below the SD line, most are shorter!
The average height of the sons of 72” fathers is given by the graph of averages, which is approximated by the regression line:
\[
\hat{y} = 69 + 0.5 \times \frac{2.7}{2.7} \times (72 - 68) = 69 + 2 = 71
\]

Thus the sons are on average shorter than the fathers. Similarly, consider 64” fathers – 4” below average.
The same calculation says that their sons on average are 2 inches below average, i.e. 69 + 2 = 67 inches.
Tall fathers have mostly shorter sons/short fathers mostly taller sons. [The aristocratic Galton called this “regression to mediocrity”].

This phenomenon is general: it applies to football-shaped point clouds (bivariate normal): to test scores, to the stock market etc.

4. THE CHANCE ERROR MODEL Return to the two quizzes example. This is another example of the regression effect. It is a consequence of correlation \( r \) being less than 1. Indeed, the regression fallacy occurs when you argue that there is some substantive reason other than chance variation going on.

Here is another way to think about the regression effect. In the test-retest situation, we make a model that

\[
Y = T + e \quad \text{(test score = true score + chance error)}
\]

Assume the true scores follow the normal density curve, mean 100, SD 15. \( T \sim N(100, 15) \)
Suppose the chance error is as likely to be positive as negative, and is around 5 in size: \( e \sim (0, 5) \)
(For simplicity, could imagine that \( e \) is either +5 or −5 with 50-50 chance.) Take people who scored 140 on the test. Two possibilities:

- a) true score below 140, positive chance error (\( T < 140, + \) error) e.g. 135+5
- b) true score above 140, negative chance error (\( T > 140, - \) error) e.g. 145-5

A plot of the normal curve shows that the first explanation is more likely – the true score is most likely lower, and so on average, the scores on the second test will be a bit lower than the first.

Population definition of correlation, for jointly distributed random variables \((X,Y)\):

\[
\rho = \frac{\text{Cov}(X,Y)}{\sqrt{\text{Var} X \cdot \text{Var} Y}} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}
\]

where \( \text{Cov}(X,Y) = E[(X - \mu_X)(Y - \mu_Y)] \).
If \( X \) and \( Y \) are independent, then \( \text{Cov}(X,Y) = 0 \) and \( \rho = 0 \).
(Proof: \( \text{Cov}(X,Y) = E(X - \mu_X)(Y - \mu_Y) = E(X - \mu_X) \cdot E(Y - \mu_Y) \) using independence = 0]
Population version of least squares:
We seek \( \alpha \) and \( \beta \) to minimize \((a,b) \rightarrow E[Y - a - bX]^2\)
To minimize, set partial derivatives \( \frac{\partial}{\partial a} \) and \( \frac{\partial}{\partial b} \) equal to zero:
\[
\frac{\partial}{\partial a} = 0 \Rightarrow -2E[Y - \alpha - \beta X] = 0 \Rightarrow \mu_Y = \alpha + \beta \mu_x
\]
\[ \frac{\partial}{\partial b} = 0 \Rightarrow -2E[Y - \alpha - \beta X] = 0 \]

\[ \Rightarrow EXY = \alpha \mu_x + \beta E X^2 \]

\[ = \mu_y \mu_x + \beta [E X^2 - \mu_x^2] \]

\[ \Rightarrow \beta \text{Var} X = E(X - \mu_X)(Y - \mu_Y) = \text{Cov}(X, Y) \]

\[ \Rightarrow \beta = \frac{\sigma_{XY}}{\sigma_X^2} = \rho \frac{\sigma_X \sigma_Y}{\sigma_X^2} = \rho \frac{\sigma_Y}{\sigma_X} \]