Regression and Correlation

Fathers’ and daughters’ heights

\[ \text{corr} = 0.52 \]
Linear regression

![Graph showing the relationship between father's height and daughter's height.](image)

Father's height (inches)

Daughter's height (inches)
Regression line

\[ \text{Slope} = r \times \frac{\text{SD}(Y)}{\text{SD}(X)} \]

SD line

\[ \text{Slope} = \frac{\text{SD}(Y)}{\text{SD}(X)} \]
SD line vs regression line

Both lines go through the point \((\bar{X}, \bar{Y})\).

Predicting father’s ht from daughter’s ht
Predicting father’s ht from daughter’s ht

![Scatter plot showing height relationship between fathers and daughters.]

Predicting father’s ht from daughter’s ht

![Scatter plot showing height relationship between fathers and daughters with a trend line.]

There are two regression lines!

The equations

Regression of $y$ on $x$  (for predicting $y$ from $x$)

Slope = $r \frac{SD(y)}{SD(x)}$  \hspace{1cm}  \text{Goes through the point } (\bar{x}, \bar{y})

$\hat{y} - \bar{y} = r \frac{SD(y)}{SD(x)} (x - \bar{x})$

$\rightarrow \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$  \hspace{1cm}  \text{where } \hat{\beta}_1 = r \frac{SD(y)}{SD(x)} \text{ and } \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

Regression of $x$ on $y$  (for predicting $x$ from $y$)

Slope = $r \frac{SD(x)}{SD(y)}$  \hspace{1cm}  \text{Goes through the point } (\bar{y}, \bar{x})

$\hat{x} - \bar{x} = r \frac{SD(x)}{SD(y)} (y - \bar{y})$

$\rightarrow \hat{x} = \hat{\beta}^*_0 + \hat{\beta}^*_1 y$  \hspace{1cm}  \text{where } \hat{\beta}^*_1 = r \frac{SD(x)}{SD(y)} \text{ and } \hat{\beta}^*_0 = \bar{x} - \hat{\beta}^*_1 \bar{y}$
Span and height

Variability

Having no information about $x$,

Predict $y$ as $\bar{y}$

Typical prediction error: $SD(y)$

For predicting height, $SD(y) \approx 2.73$

Having been told about $x$,

Predict $y$ using the regression line: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

Typical prediction error: $SD(y) \sqrt{1 - r^2}$

For predicting height from span, $SD(y) \sqrt{1 - r^2} \approx 1.71$
Correlation vs Regression

- In a correlation setting we try to determine whether two random variables vary together (covary).

- There is no ordering between those variables, and we do not try to explain one of the variables as a function of the other.

- In regression settings we describe the dependence of one variable on the other variable.

- There is an ordering of the variables, often called the dependent variable and the independent variable.

The correlation coefficient of two jointly distributed random variables $X$ and $Y$ is defined as

$$
\rho = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}
$$

where $\text{cov}(X, Y)$ is the covariance between $X$ and $Y$, and $\sigma_X$ and $\sigma_Y$ are their respective standard deviations.

If $X$ and $Y$ follow a bivariate normal distribution with correlation $\rho$

$$
\begin{pmatrix}
  x_i \\
  y_i
\end{pmatrix}
\sim N
\left( \begin{pmatrix}
  \mu_X \\
  \mu_Y
\end{pmatrix},
\begin{pmatrix}
  \sigma_X^2 & \rho \sigma_X \sigma_Y \\
  \rho \sigma_X \sigma_Y & \sigma_Y^2
\end{pmatrix}\right)
$$

then

$$
y_i|x_i \sim N \left( \beta_0 + \beta_1 x_i, \sigma^2 \right)
$$

where $\beta_0 = \mu_Y - \beta_1 \mu_X$, $\beta_1 = \rho \sigma_Y / \sigma_X$, and $\sigma^2 = \sigma_Y^2(1 - \rho^2)$. 
Correlation?

$R^2 = 0.56$

$R^2 = 0.85$

Gene expression
Concordance

If you want to show that two sets of measurements are alike (such as gene expression from two technical replicates of the same sample) use the concordance correlation coefficient.

The concordance correlation between two random variables $X$ and $Y$ is defined as

$$
\rho_{CC}(X, Y) = \frac{2 \times \text{cov}(X, Y)}{\sigma_X^2 + \sigma_Y^2 + (\mu_X - \mu_Y)^2}.
$$

Unlike the Pearson correlation coefficient, the concordance correlation is not invariant to changes in location and scale, and assesses the actual agreement between $X$ and $Y$, rather than their correlation alone.

Correlation..?