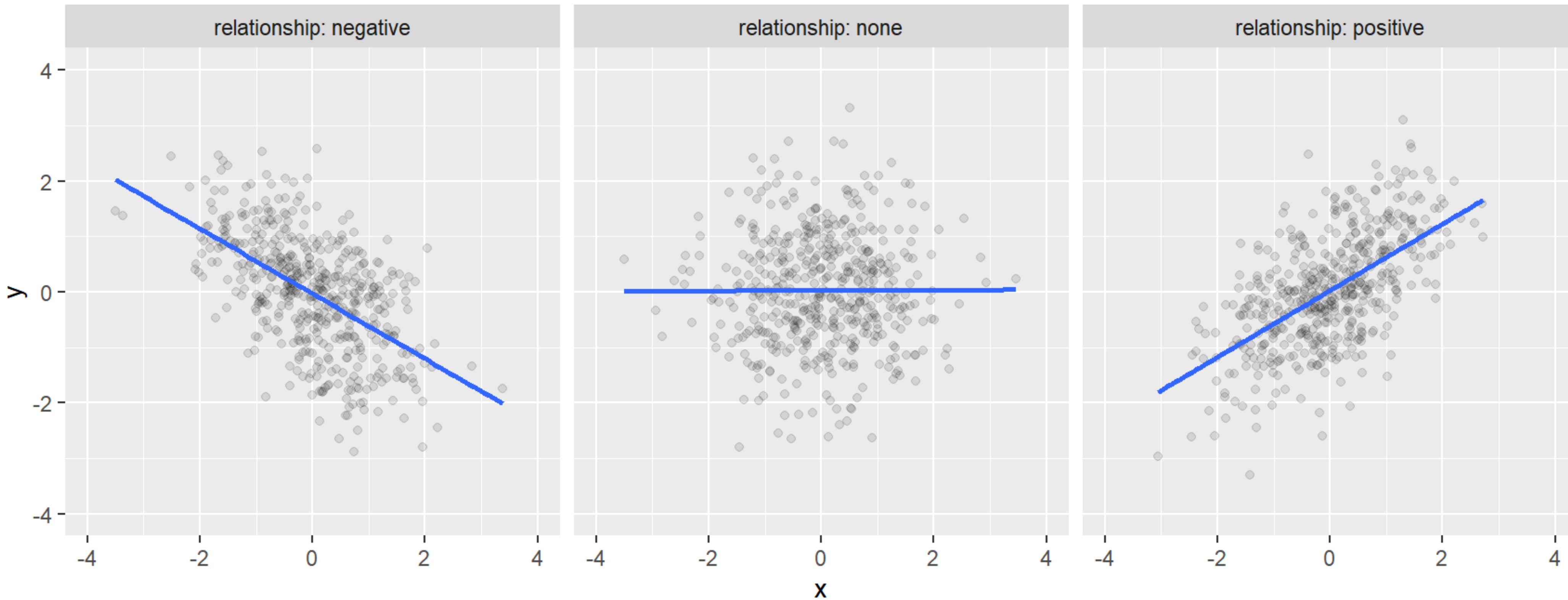
Statistical Models

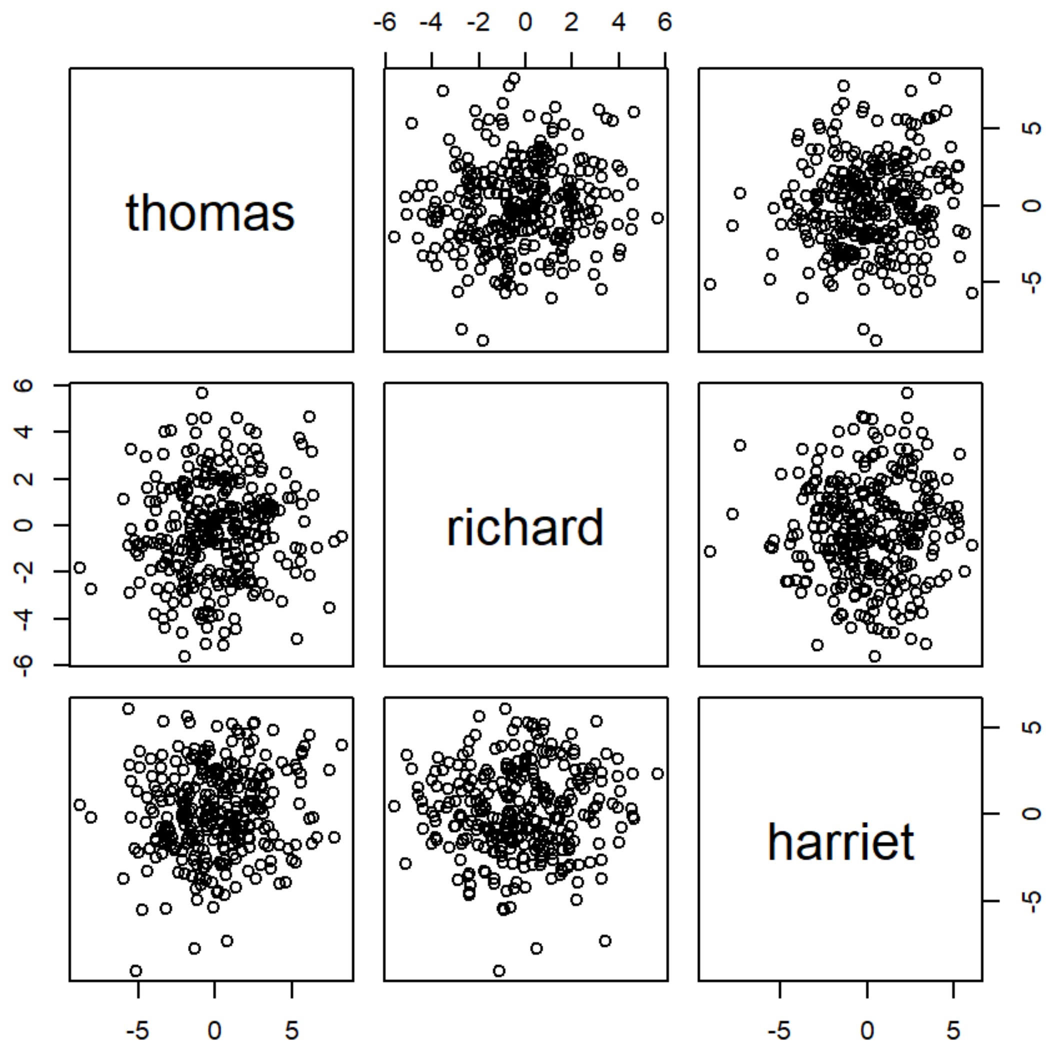
Dale Barr

University of Glasgow

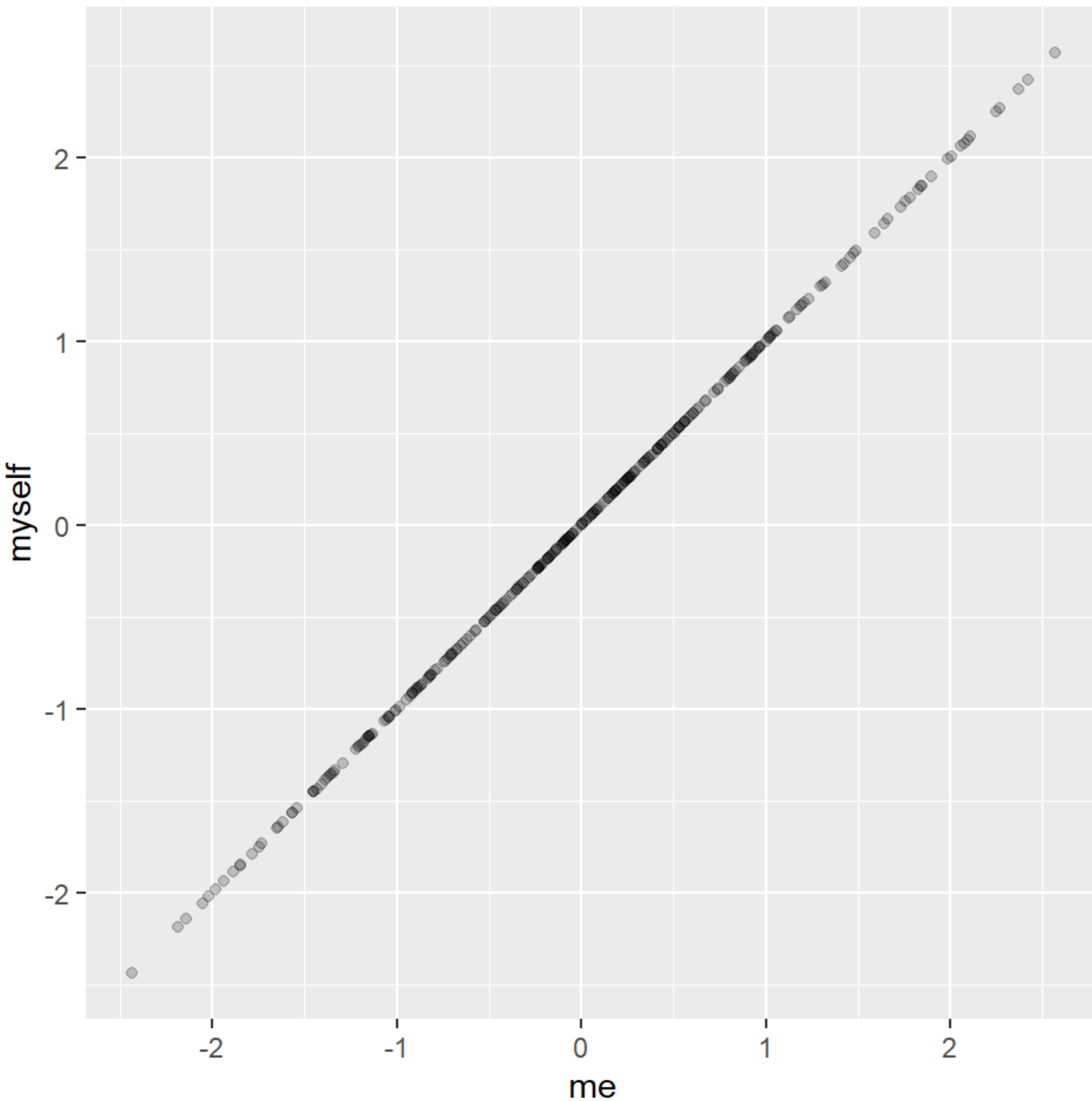
relationships



multiple relationships



the perfect relationship



today’s lecture

- correlations and correlation matrices
- simulating bivariate data
- relationship between correlation and regression

correlation coefficient

Typicaly denoted as ρ (Greek symbol 'rho') or r

$-1 \geq r \leq 1$

- $r > 0$: positive relationship
- $r < 0$: negative relationship
- $r = 0$: no relationship

Estimated using Pearson or Spearman (rank) method

c- `cor()`, `cor.test()`, `corrr::correlate()`

assumptions

- relationship between X and Y is *linear*
- deviations from line of best fit are *normally distributed*

multiple correlations

For n variables, you have

$$\frac{n!}{2(n-2)!}$$

unique pairwise relationships, where (n!) is the *factorial* of (n).

`choose(n, 2)`

```r
choose(6, 2)
```

```
[1] 15
```

```r
choose(8, 2)
```

```
[1] 28
```

correlation matrices

	IQ	verbal fluency	digit span
IQ	1.00	0.56	0.43
verbal fluency	0.56	1.00	-0.23
digit span	0.43	-0.23	1.00

`corrr::correlate()`

covariance matrices

- covariance(X,Y): $\rho_{xy}\sigma_x\sigma_y$
- covariance(X,X): $\rho_{xx}\sigma_x\sigma_x = \sigma^2$

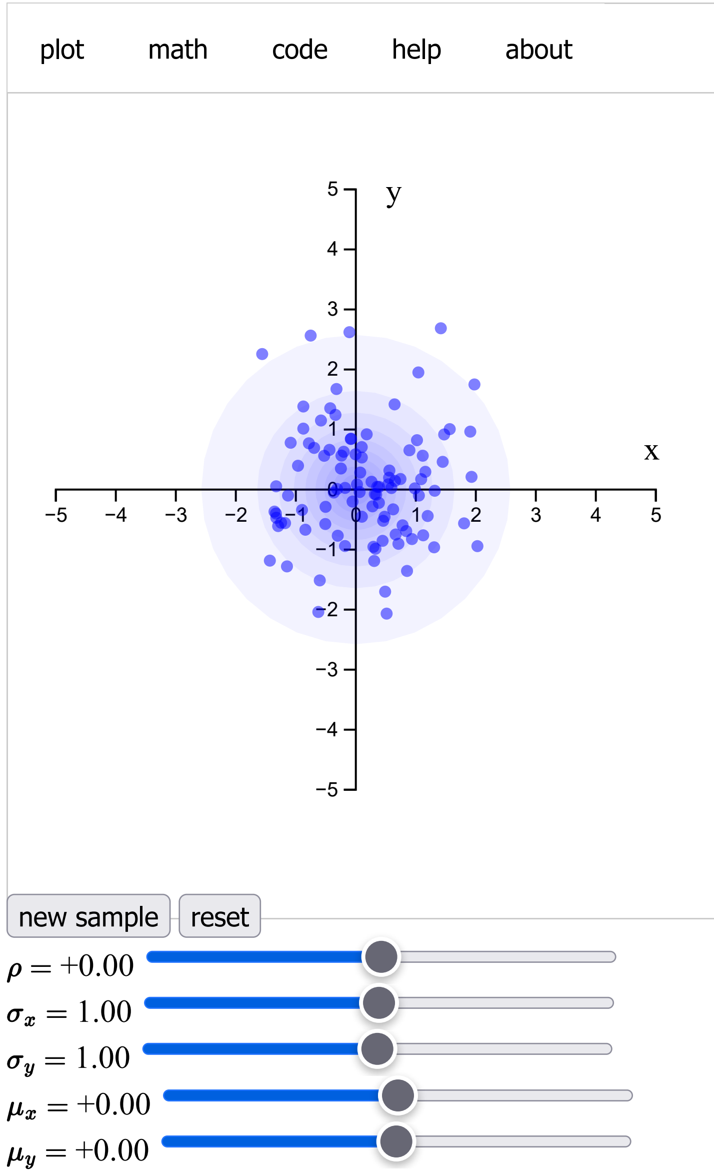
ρ_{xy}: correlation between x, y; σ_x: sd of x

A matrix that characterizes the spread of multivariate values.

$$\begin{pmatrix} {\sigma_x}^2 & \rho_{xy}\sigma_x\sigma_y \\ \rho_{yx}\sigma_y\sigma_x & {\sigma_y}^2 \end{pmatrix}$$

Usually denoted by $\mathbf{\Sigma}$; Must be *symmetric* and *positive definite*

bivariate distribution



4x4 matrix

A 4x4 covariance matrix with variables W, X, Y, Z.

$$\begin{pmatrix} \rho_{ww}\sigma_w\sigma_w & \rho_{wx}\sigma_w\sigma_x & \rho_{wy}\sigma_w\sigma_y & \rho_{wz}\sigma_w\sigma_z \\ \rho_{xw}\sigma_x\sigma_w & \rho_{xx}\sigma_x\sigma_x & \rho_{xy}\sigma_x\sigma_y & \rho_{xz}\sigma_x\sigma_z \\ \rho_{yw}\sigma_y\sigma_w & \rho_{yx}\sigma_y\sigma_x & \rho_{yy}\sigma_y\sigma_y & \rho_{yz}\sigma_y\sigma_z \\ \rho_{zw}\sigma_z\sigma_w & \rho_{zx}\sigma_z\sigma_x & \rho_{zy}\sigma_z\sigma_y & \rho_{zz}\sigma_z\sigma_z \end{pmatrix}$$

4x4 matrix

A 4x4 covariance matrix with variables W, X, Y, Z.

$$\begin{pmatrix} {\sigma_w}^2 & \rho_{wx}\sigma_w\sigma_x & \rho_{wy}\sigma_w\sigma_y & \rho_{wz}\sigma_w\sigma_z \\ \rho_{xw}\sigma_x\sigma_w & {\sigma_x}^2 & \rho_{xy}\sigma_x\sigma_y & \rho_{xz}\sigma_x\sigma_z \\ \rho_{yw}\sigma_y\sigma_w & \rho_{yx}\sigma_y\sigma_x & {\sigma_y}^2 & \rho_{yz}\sigma_y\sigma_z \\ \rho_{zw}\sigma_z\sigma_w & \rho_{zx}\sigma_z\sigma_x & \rho_{zy}\sigma_z\sigma_y & {\sigma_z}^2 \end{pmatrix}$$

diagonal matrix

$$\begin{pmatrix} {\sigma_w}^2 & 0 & 0 & 0 \\ 0 & {\sigma_x}^2 & 0 & 0 \\ 0 & 0 & {\sigma_y}^2 & 0 \\ 0 & 0 & 0 & {\sigma_z}^2 \end{pmatrix}$$

simulating correlated data

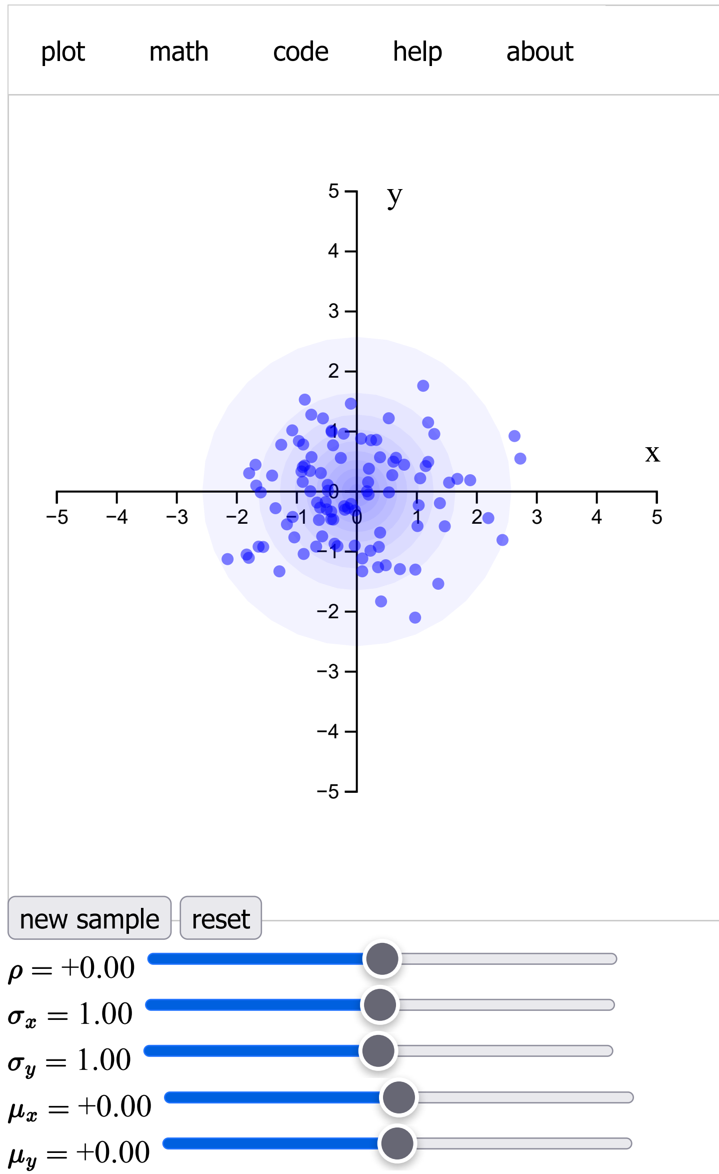
To simulate bivariate (or multivariate) data in R, use `MASS::mvrnorm()`.

`mvrnorm(n, mu, Sigma, ...)`

You need the following information:

- means of X and Y, $\bar{X}$ and $\bar{Y}$
- standard deviations of X and Y, σ_X and σ_Y.
- correlation coefficient ρ_{XY}.

simulating bivariate data



correlation and the GLM

$$Y_i = \beta_0 + \beta_1 X_i + e_i$$

$$e_i \sim N\left(0, \sigma^2\right)$$

$$\beta_1 = \rho_{XY} \frac{\sigma_Y}{\sigma_X}$$

$$\beta_0 = \bar{Y} - \beta_1 \bar{X}$$