CORRELATION AND REGRESSION

STATISTICAL MODELS DALE BARR

PSYCHOLOGY, UNIVERSITY OF GLASGOW

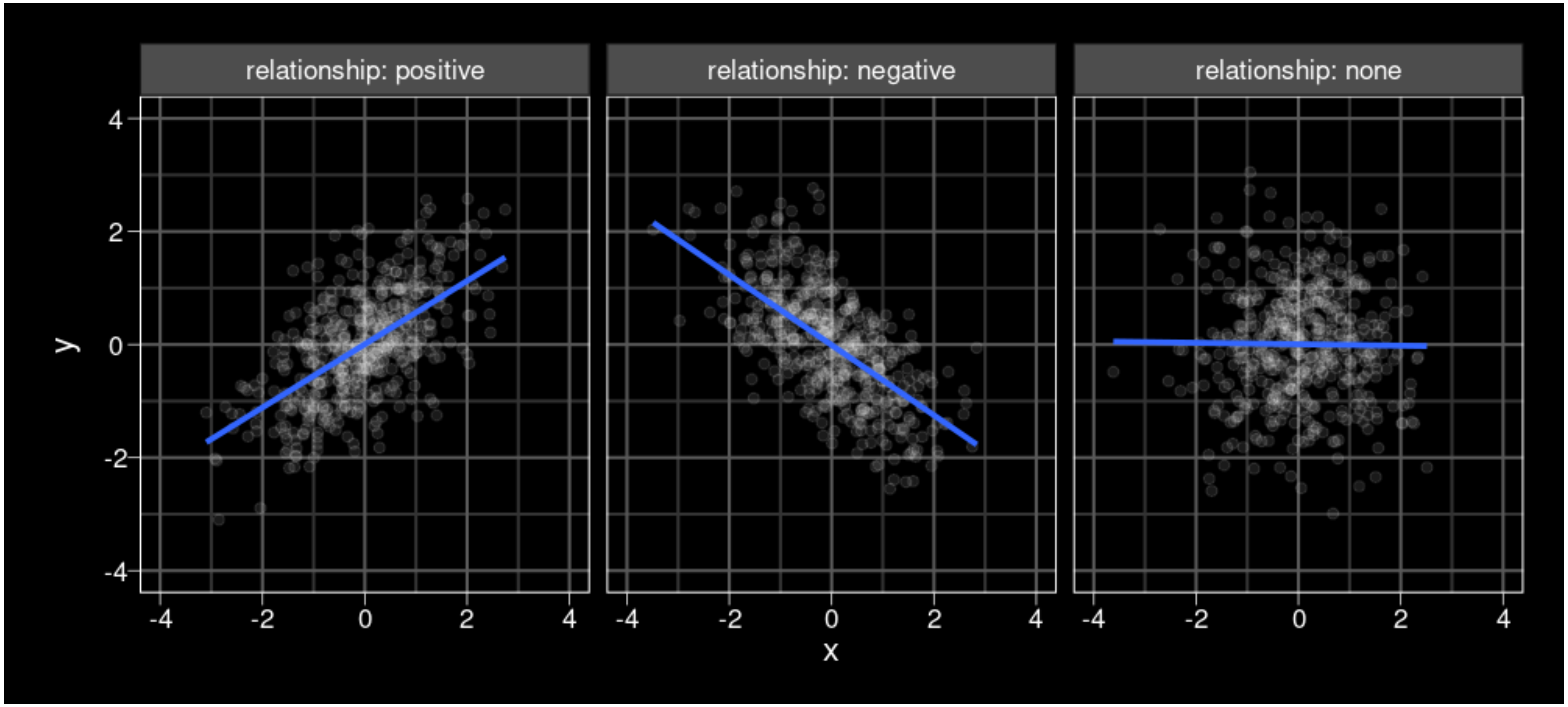
Created: 2020-10-05 Mon 09:34

TODAY'S LECTURE

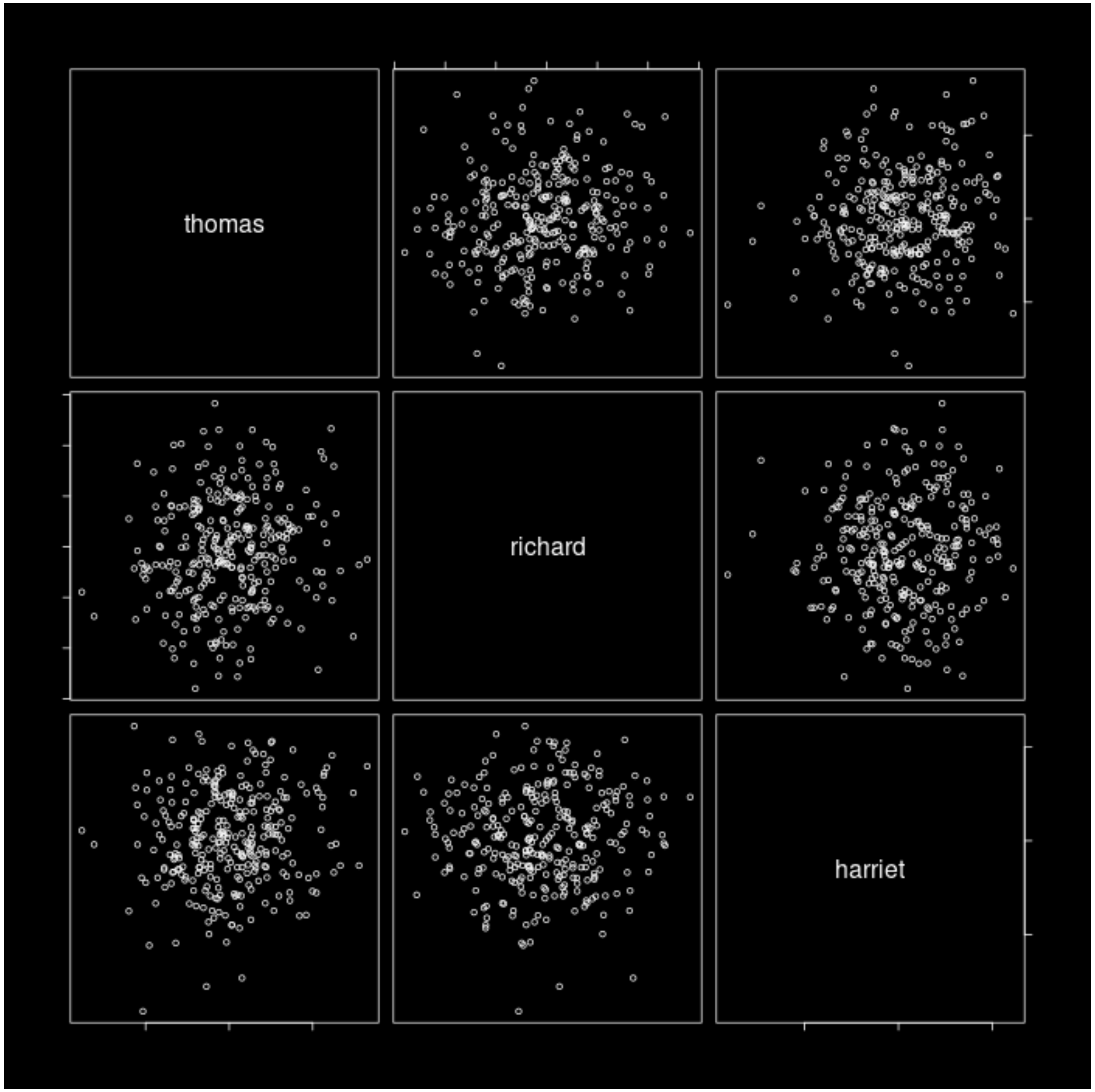
1. correlations and correlation matrices
2. simulating correlational data
3. relationship between correlation and regression

CORRELATIONS

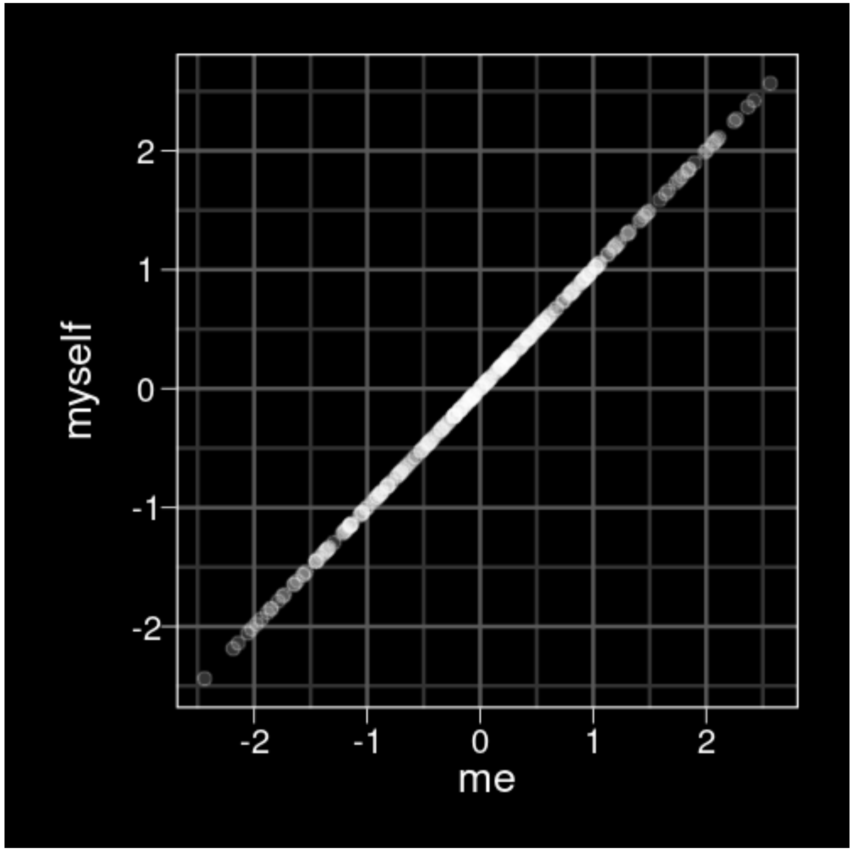
RELATIONSHIPS

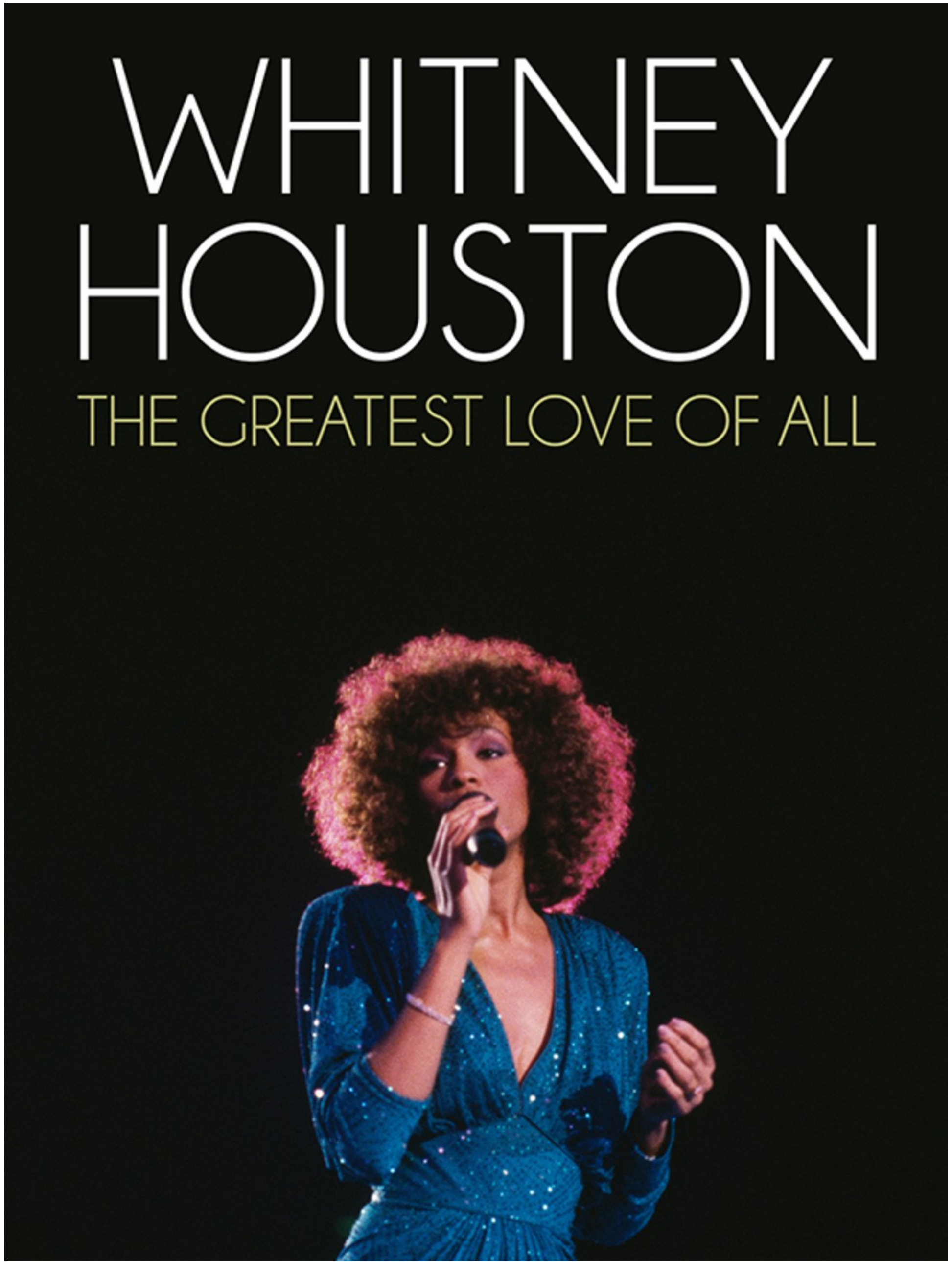


MULTIPLE RELATIONSHIPS



THE PERFECT RELATIONSHIP





THE 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

r^2: *coefficient of determination* (shared variance)

Estimated using Pearson or Spearman (rank) method. In R: `cor()`, `cor.test()`

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.

In R: `choose(n, 2)`.

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

In R: `corrr::correlate()`

CORRELATION MATRICES

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

SIMULATING CORRELATIONAL 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, μ_x and μ_y
- standard deviations of X and Y, σ_x and σ_y.
- correlation coefficient ρ_{xy}.

THE `bivariate` APP

https://shiny.psy.gla.ac.uk/Dale/bivariate

REVIEW: STANDARD DEVIATION

a measure of how much some quantity varies

"standard deviation of x": σ_x

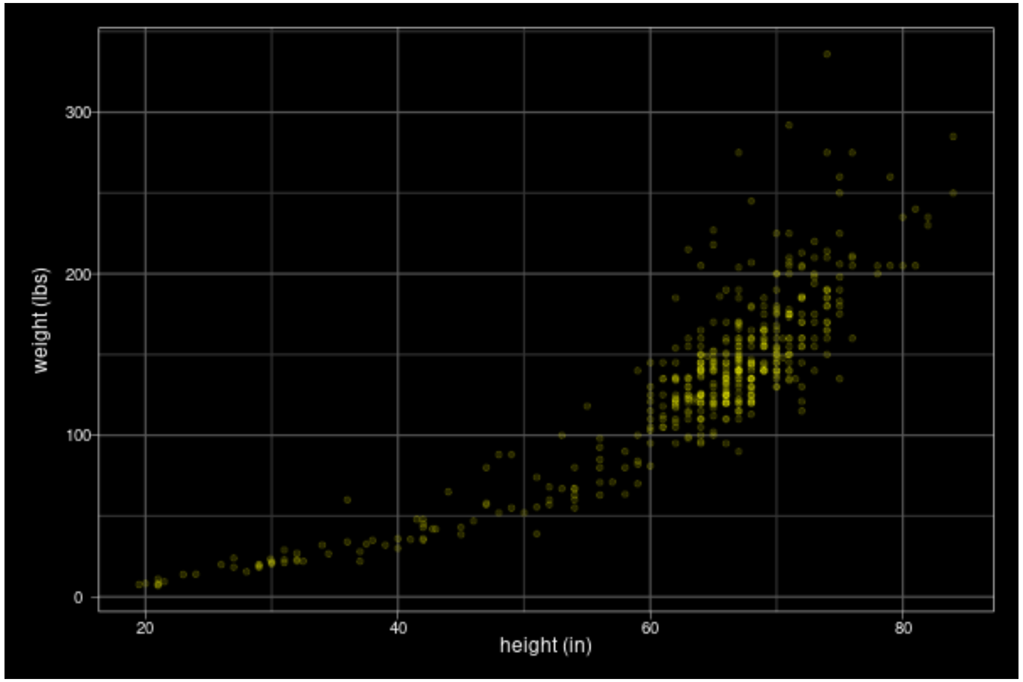
"variance of x": ${\sigma_x}^2$

- estimating σ_x from a sample:

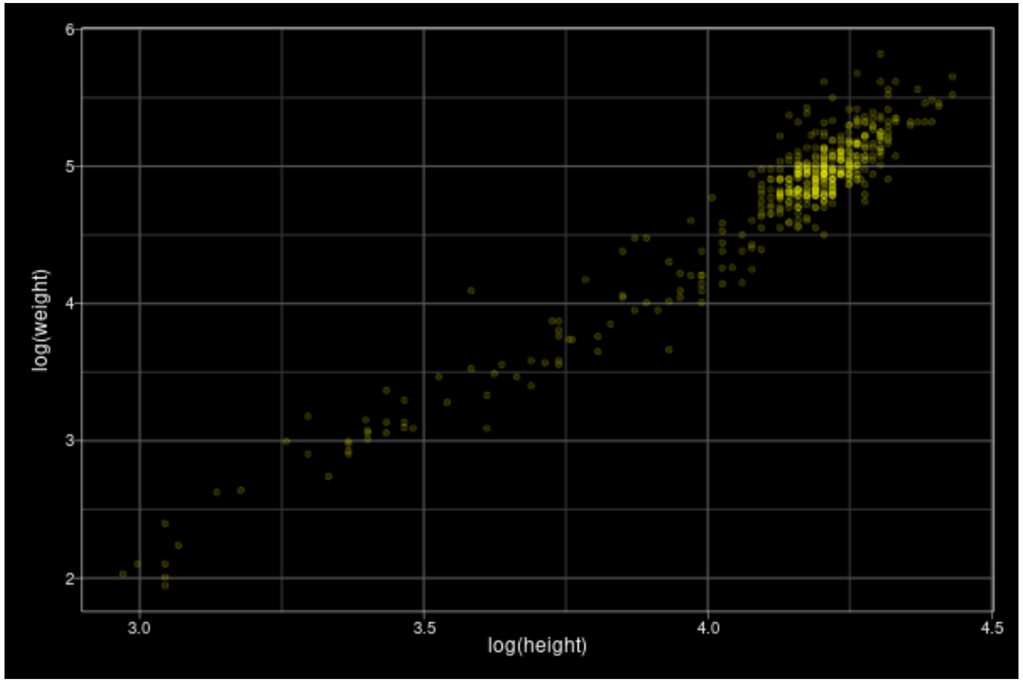
$$\hat{\sigma}_x = \sqrt{\frac{\Sigma(X - \hat{\mu}_x)^2}{N-1}}$$

LET'S MAKE SYNTHETIC HUMANS

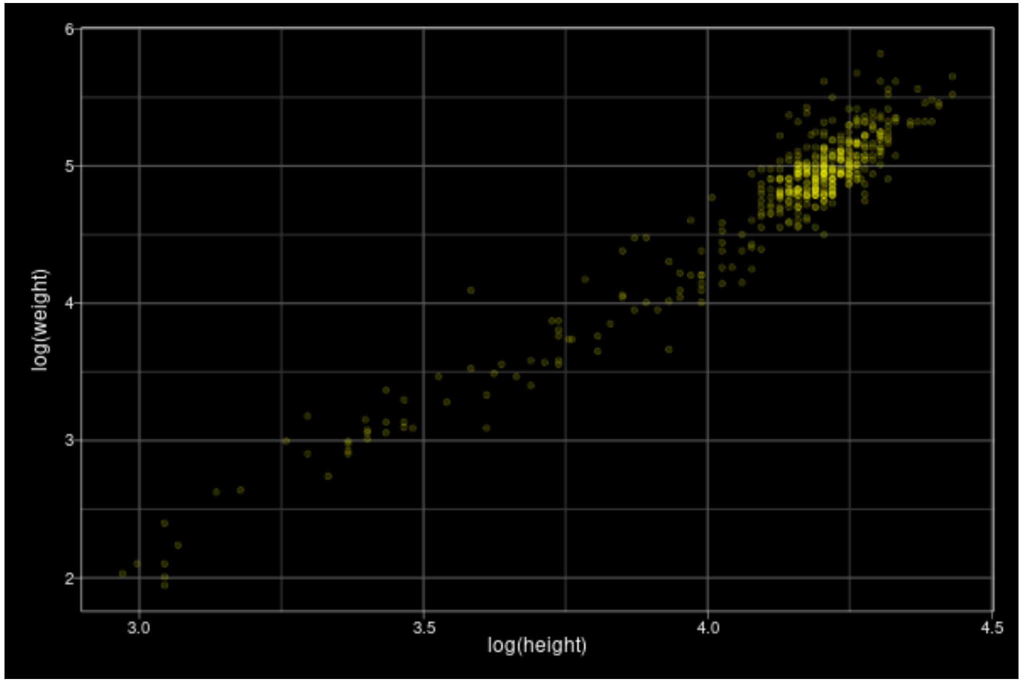
height and weight measurements for 435 people, taken from here



LOG-TRANSFORMED DATA



SUMMARY STATISTICS



$\hat{\mu}_x$	4.11
$\hat{\mu}_y$	4.74
$\hat{\sigma}_x$	.26
$\hat{\sigma}_y$	.65
$\hat{\rho}_{xy}$	.96

COVARIANCE MATRIX

$$\boldsymbol{\Sigma}$$

A square matrix that characterizes the variances and their interrelationships (covariances).

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

Must be **symmetric** and **positive definite**

CALCULATIONS

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

σ_x	.26
σ_y	.65
ρ_{xy}	.96

SIMULATING WITH MASS::mvrnorm()

```r
my_cov <- .96 * .26 * .65
my_Sigma <- matrix(c(.26^2, my_cov,
                     my_cov, .65^2),
                   ncol = 2)
my_Sigma
```

```
        [,1]    [,2]
[1,] 0.06760 0.16224
[2,] 0.16224 0.42250
```

```r
set.seed(62)

## DON'T put library(MASS)
## in your script!
newpeeps <-
  MASS::mvrnorm(6,
                mu = c(height = 4.11,
                       weight = 4.74),
                Sigma = my_Sigma)

newpeeps
```

```
       height   weight
[1,] 4.254209 5.282913
[2,] 4.257828 4.895222
[3,] 3.722376 3.759767
[4,] 4.191287 4.764229
[5,] 4.739967 6.185191
[6,] 4.058105 4.806485
```

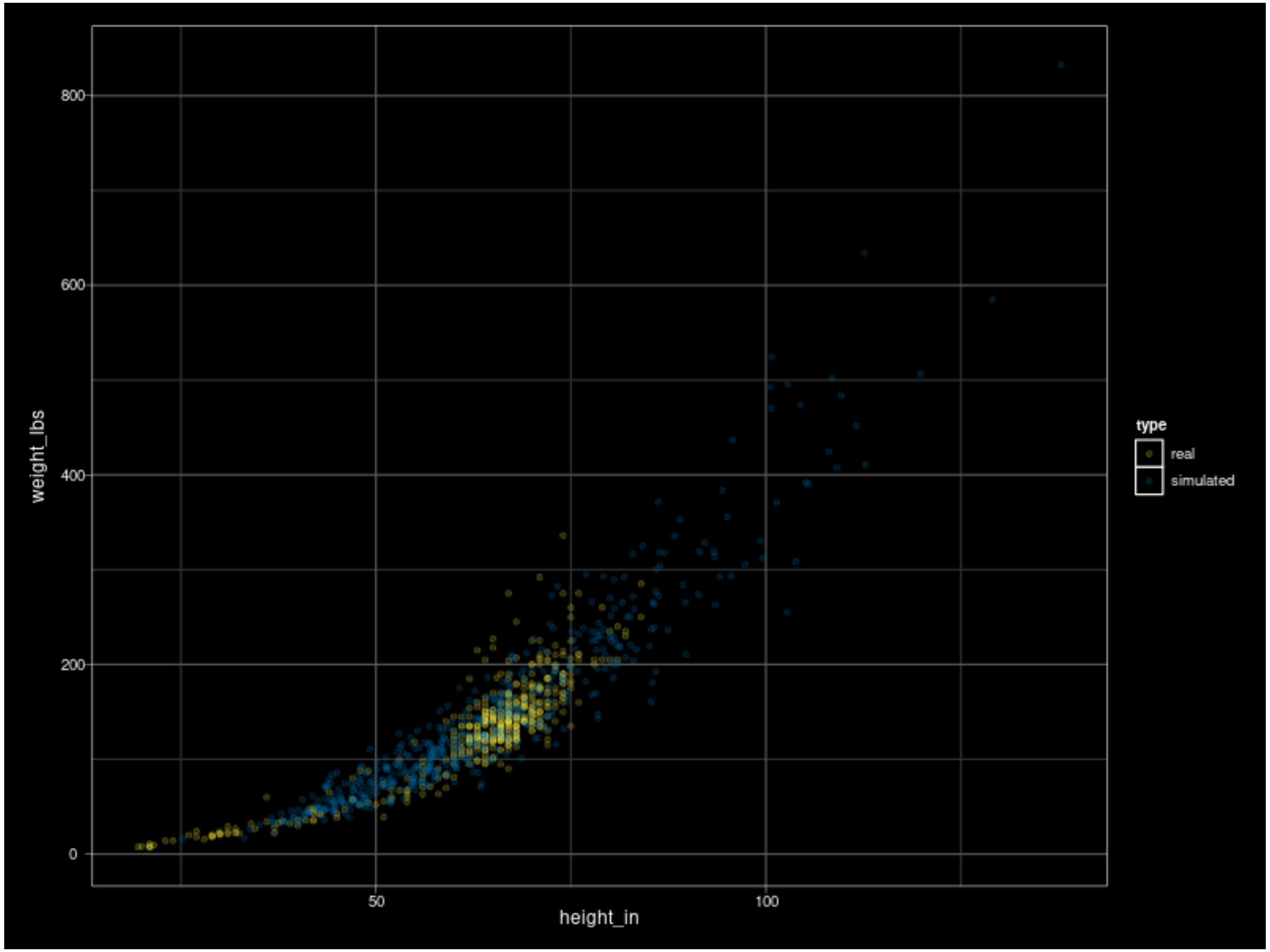
TRANSFORM BACK TO RAW UNITS

The `exp()` function is the inverse of `log()`.

```
exp(newpeeps)
```

```
        height    weight
[1,]  70.40108 196.94276
[2,]  70.65632 133.64963
[3,]  41.36254  42.93844
[4,]  66.10779 117.24065
[5,] 114.43045 485.50576
[6,]  57.86453 122.30092
```

OUR SYNTHETIC HUMANS



RELATIONSHIP BETWEEN CORRELATION AND REGRESSION

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

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

$$\beta_1 = \rho_{xy} \frac{\sigma_y}{\sigma_x}$$

$$\beta_0 = \mu_y - \beta_1 \mu_x$$

IMPLICATIONS

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

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

$$\beta_1 = \rho_{xy} \frac{\sigma_y}{\sigma_x}$$

$$\beta_0 = \mu_y - \beta_1 \mu_x$$

- $\beta_1 > 0$ implies $\rho > 0$, since standard deviations can't be negative.
- $\beta_1 < 0$ implies $\rho < 0$, for the same reason.
- Rejecting $H_0 : \beta_1 = 0$ is the same as rejecting $H_0 : \rho = 0$.
  - also, same p-values for β_1 in `lm()` as for r in `cor.test()`.

REGRESSION FROM CORRELATION

A study of student performance obtains a correlation of .16 between final exam score and number of lectures attended. The mean score on the final exam was 70 ($SD = 10$), and the mean number of courses attended was 6 ($SD = 2$).

Write the regression equation predicting exam score from attendance.

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

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

$$\beta_1 = \rho_{xy} \frac{\sigma_y}{\sigma_x}$$

$$\beta_0 = \mu_y - \beta_1 \mu_x$$

CORRELATION FROM REGRESSION

A study on the relationship between wellbeing and hours spent on social media (per week) yields the following regression:

wellbeing = 62 - .3 hours

with 5 for the standard deviation of wellbeing and .1 for the standard deviation of number of hours.

What is the correlation?

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

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

$$\beta_1 = \rho_{xy} \frac{\sigma_y}{\sigma_x}$$

$$\beta_0 = \mu_y - \beta_1 \mu_x$$

Created by Dale Barr