Learning Objectives

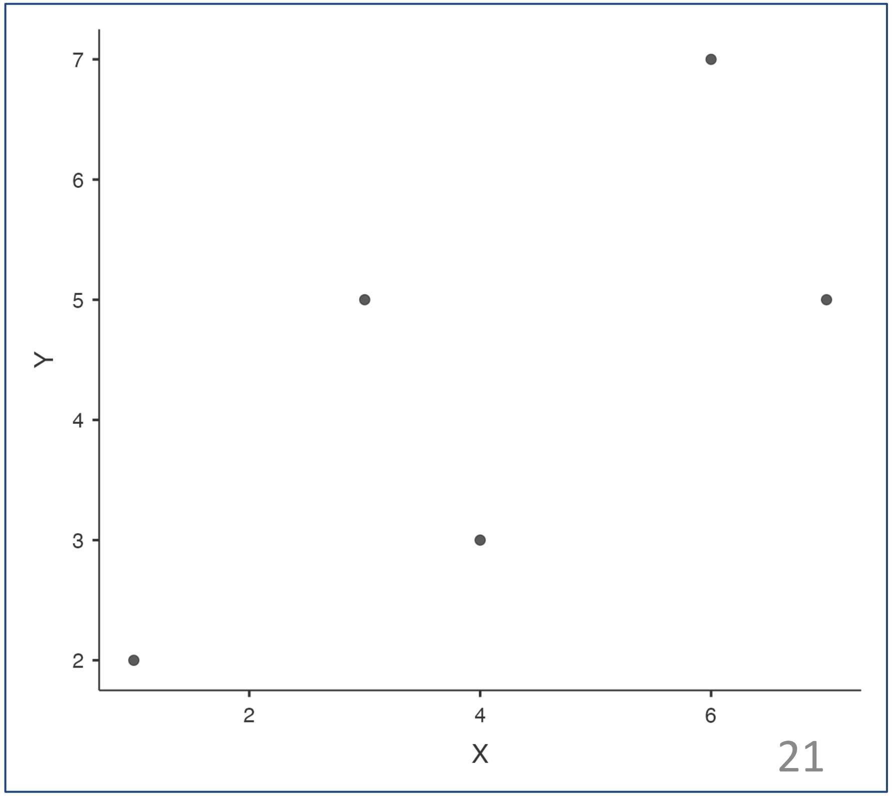
- **Describe** *correlation* in terms of direction, magnitude, and form
- **Build** intuitions about correlations based on visual *scatterplots*
- **Calculate** *correlation coefficient* and the *coefficient of determination* given sets of data
- **Describe** other correlation statistics and when we might use them

Example pp. 134

$$\text{Pearson's } r = \frac{\sum XY - \frac{(\sum X)(\sum Y)}{N}}{\sqrt{\left[\sum X^2 - \frac{(\sum X)^2}{N}\right]\left[\sum Y^2 - \frac{(\sum Y)^2}{N}\right]}}$$

ID	*X*	*Y*
A	1	2
B	3	5
C	4	3
D	6	7
E	7	5

Value of *X* plotted against value of *Y*

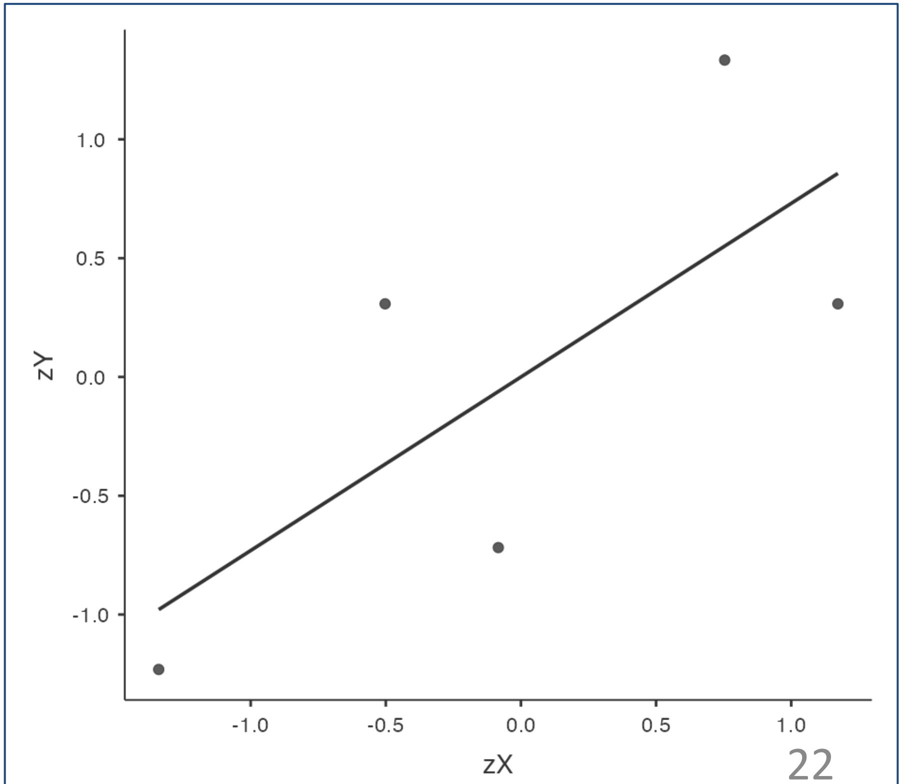


Example pp. 134

$$\text{Pearson's } r = \frac{\sum XY - \frac{(\sum X)(\sum Y)}{N}}{\sqrt{\left[\sum X^2 - \frac{(\sum X)^2}{N}\right]\left[\sum Y^2 - \frac{(\sum Y)^2}{N}\right]}}$$

ID	*X*	*Y*
A	1	2
B	3	5
C	4	3
D	6	7
E	7	5

Value of *zX* plotted against value of *zY*



Example pp. 134

$$\text{Pearson's } r = \frac{\sum XY - \frac{(\sum X)(\sum Y)}{N}}{\sqrt{\left[\sum X^2 - \frac{(\sum X)^2}{N}\right]\left[\sum Y^2 - \frac{(\sum Y)^2}{N}\right]}}$$

ID	*X*	*Y*
A	1	2
B	3	5
C	4	3
D	6	7
E	7	5
N =	ΣX =	ΣY =

Example pp. 134

$$\text{Pearson's } r = \frac{\sum XY - \dfrac{(\sum X)(\sum Y)}{N}}{\sqrt{\left[\sum X^2 - \dfrac{(\sum X)^2}{N}\right]\left[\sum Y^2 - \dfrac{(\sum Y)^2}{N}\right]}}$$

ID	X	Y	X^2	Y^2	XY
A	1	2	1	4	2
B	3	5	9	25	15
C	4	3	16	9	12
D	6	7	36	49	42
E	7	5	49	25	35
$N = 6$	$\Sigma X = 21$	$\Sigma Y = 22$	$\Sigma X^2 = 111$	$\Sigma Y^2 = 112$	$\Sigma XY = 106$

Calculating Pearson's *r*

Method 2: Raw Scores

$$\text{Pearson's } r = \frac{\sum XY - \frac{(\sum X)(\sum Y)}{N}}{\sqrt{\left[\sum X^2 - \frac{(\sum X)^2}{N}\right]\left[\sum Y^2 - \frac{(\sum Y)^2}{N}\right]}}$$

$N = 5$

$\Sigma X = 21$

$\Sigma Y = 22$

$(\Sigma X)^2 =$

$(\Sigma Y)^2 =$

$\Sigma X^2 = 111$

$\Sigma Y^2 = 112$

$\Sigma XY = 106$

$$\frac{106 - \frac{(21)(22)}{5}}{}$$

Calculating Pearson's *r*

Method 2: Raw Scores

$$\text{Pearson's } r = \frac{\sum XY - \frac{(\sum X)(\sum Y)}{N}}{\sqrt{\left[\sum X^2 - \frac{(\sum X)^2}{N}\right]\left[\sum Y^2 - \frac{(\sum Y)^2}{N}\right]}}$$

$N = 5$

$\Sigma X = 21$

$\Sigma Y = 22$

$(\Sigma X)^2 =$

$(\Sigma Y)^2 =$

$\Sigma X^2 = 111$

$\Sigma Y^2 = 112$

$\Sigma XY = 106$

$$\frac{106 - \frac{(21)(22)}{5}}{\sqrt{([111 - \frac{???}{5}][112 - \frac{???}{5}])}}$$

Calculating Pearson's *r*

Method 2: Raw Scores

$$\text{Pearson's } r = \frac{\Sigma XY - \frac{(\Sigma X)(\Sigma Y)}{N}}{\sqrt{\left[\Sigma X^2 - \frac{(\Sigma X)^2}{N}\right]\left[\Sigma Y^2 - \frac{(\Sigma Y)^2}{N}\right]}}$$

$N = 5$

$\Sigma X = 21$

$\Sigma Y = 22$

$(\Sigma X)^2 = 441$

$(\Sigma Y)^2 = 484$

$\Sigma X^2 = 111$

$\Sigma Y^2 = 112$

$\Sigma XY = 106$

$$\frac{106 - \frac{(21)(22)}{5}}{\sqrt{([111 - \frac{441}{5}][112 - \frac{484}{5}])}}$$

Calculating Pearson's *r*

Method 2: Raw Scores

$$\text{Pearson's } r = \frac{\Sigma XY - \frac{(\Sigma X)(\Sigma Y)}{N}}{\sqrt{\left[\Sigma X^2 - \frac{(\Sigma X)^2}{N}\right]\left[\Sigma Y^2 - \frac{(\Sigma Y)^2}{N}\right]}}$$

$N = 5$

$\Sigma X = 21$

$\Sigma Y = 22$

$(\Sigma X)^2 = 441$

$(\Sigma Y)^2 = 484$

$\Sigma X^2 = 111$

$\Sigma Y^2 = 112$

$\Sigma XY = 106$

$$\frac{106 - \frac{462}{5}}{\sqrt{\left(\left[111 - \frac{441}{5}\right]\left[112 - \frac{484}{5}\right]\right)}}$$

Calculating Pearson's *r*

Method 2: Raw Scores

$$\text{Pearson's } r = \frac{\sum XY - \frac{(\sum X)(\sum Y)}{N}}{\sqrt{\left[\sum X^2 - \frac{(\sum X)^2}{N}\right]\left[\sum Y^2 - \frac{(\sum Y)^2}{N}\right]}}$$

$N = 5$

$\Sigma X = 21$

$\Sigma Y = 22$

$(\Sigma X)^2 = 441$

$(\Sigma Y)^2 = 484$

$\Sigma X^2 = 111$

$\Sigma Y^2 = 112$

$\Sigma XY = 106$

$$\frac{106 - 92.4}{\sqrt{([111 - 88.2][112 - 96.8])}}$$

Calculating Pearson's r

Method 2: Raw Scores

$$\text{Pearson's } r = \frac{\Sigma XY - \frac{(\Sigma X)(\Sigma Y)}{N}}{\sqrt{\left[\Sigma X^2 - \frac{(\Sigma X)^2}{N}\right]\left[\Sigma Y^2 - \frac{(\Sigma Y)^2}{N}\right]}}$$

$N = 5$
$\Sigma X = 21$
$\Sigma Y = 22$
$(\Sigma X)^2 = 441$
$(\Sigma Y)^2 = 484$
$\Sigma X^2 = 111$
$\Sigma Y^2 = 112$
$\Sigma XY = 106$

$$\frac{13.6}{\sqrt{([22.8][15.2])}}$$

Calculating Pearson's r

Method 2: Raw Scores

$$\text{Pearson's } r = \frac{\sum XY - \frac{(\sum X)(\sum Y)}{N}}{\sqrt{\left[\sum X^2 - \frac{(\sum X)^2}{N}\right]\left[\sum Y^2 - \frac{(\sum Y)^2}{N}\right]}}$$

$N = 5$

$\Sigma X = 21$

$\Sigma Y = 22$

$(\Sigma X)^2 = 441$

$(\Sigma Y)^2 = 484$

$\Sigma X^2 = 111$

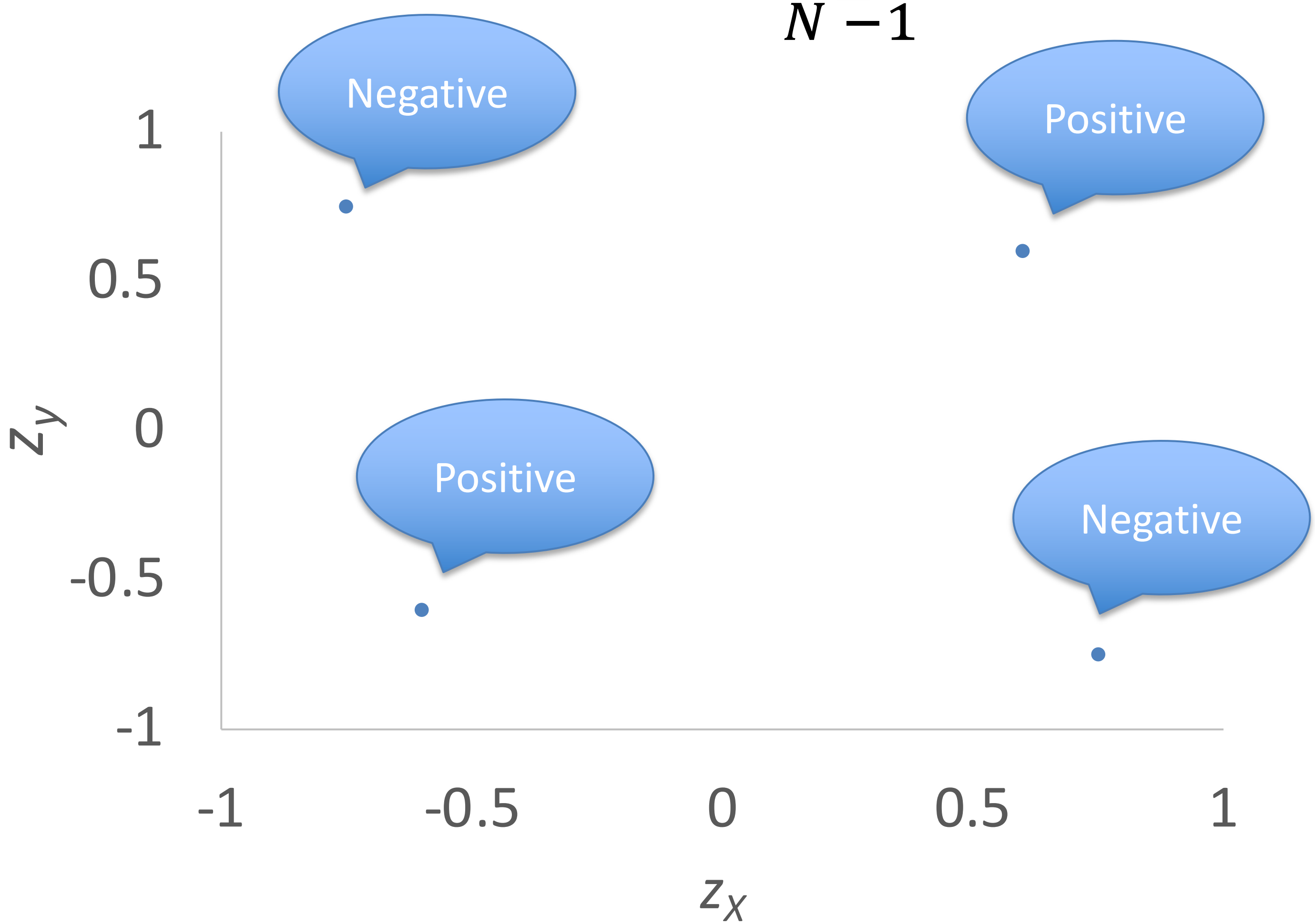
$\Sigma Y^2 = 112$

$\Sigma XY = 106$

$$\frac{13.6}{\sqrt{346.56}} = \frac{13.6}{18.616} = .7306 = r$$

Why cross-products?

- The *cross-product* is the numerator of the correlation formula: $r = \frac{\Sigma z_x z_y}{N-1}$

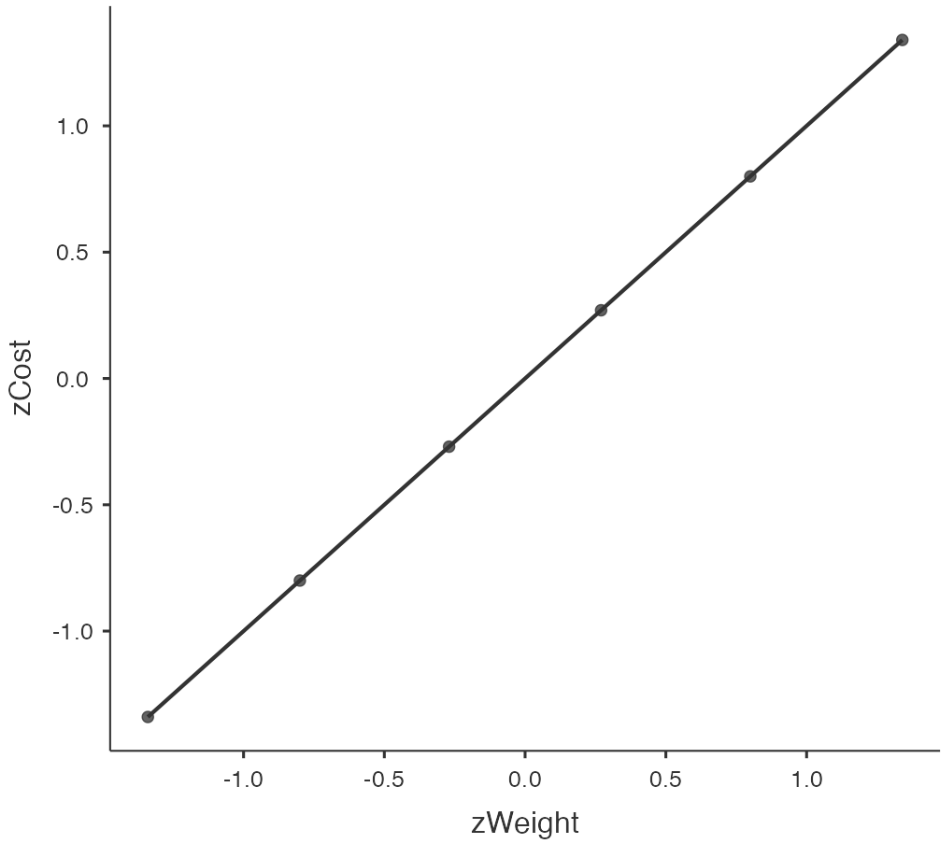


Pearson's *r*, Table 6.3

Original Data

$$\text{Pearson's } r = \frac{\Sigma z_x z_y}{N - 1}$$

ID	z_{weight}	z_{cost}	$z_x * z_y$
A	-1.34	-1.34	**POS**
B	-0.80	-0.80	**Pos**
C	-0.27	-0.27	Pos
D	0.27	0.27	Pos
E	0.80	0.80	**Pos**
F	1.34	1.34	**POS**
N = 6			Σ = **POS+**

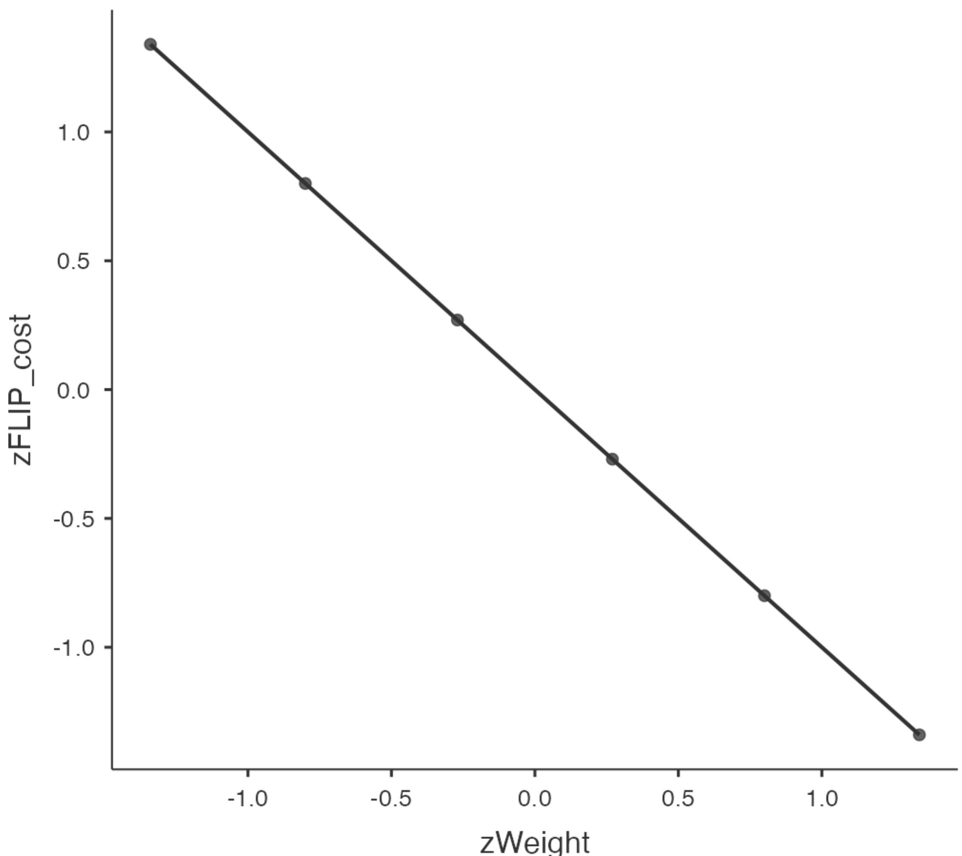


Pearson's *r*, Table 6.3

Cost variable (*Y*) has been FLIPPED

$$\text{Pearson's } r = \frac{\Sigma z_x z_y}{N-1}$$

ID	z_{weight} (*X*)	z_{cost}	$z_x * z_y$
A	-1.34	1.34	**NEG**
B	-0.80	0.80	**Neg**
C	-0.27	0.27	Neg
D	0.27	-0.27	Neg
E	0.80	-0.80	**Neg**
F	1.34	-1.34	**NEG**
N = 6			Σ = **NEG-**

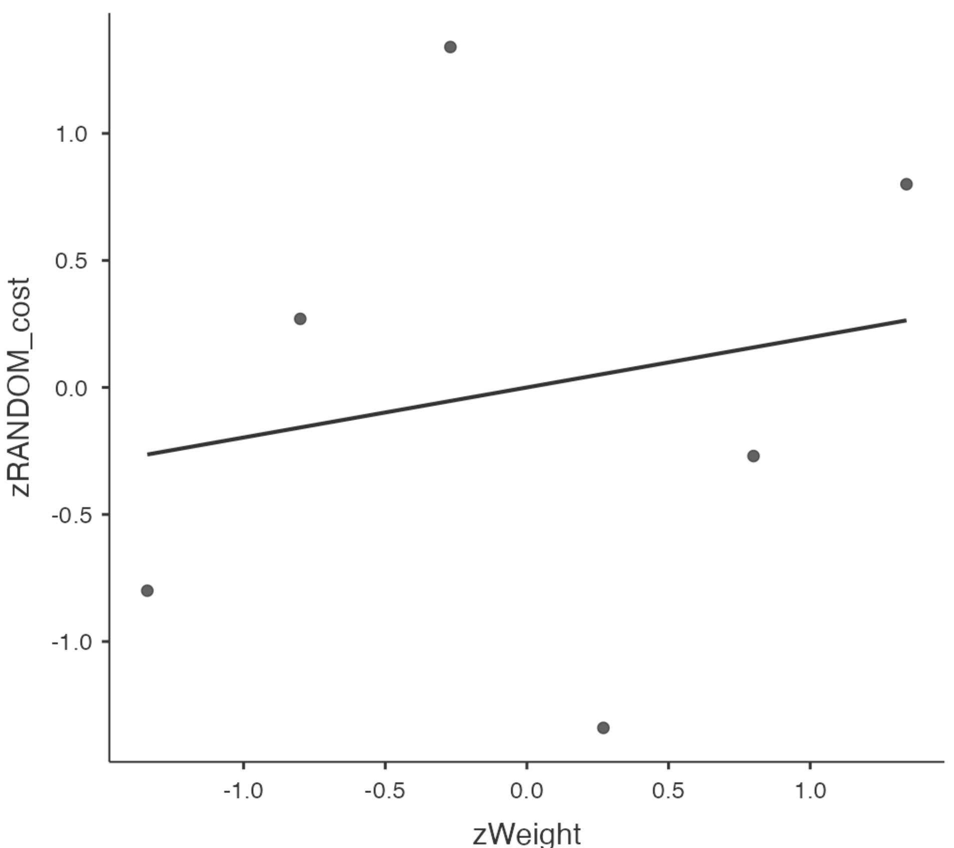


Pearson's *r*, Table 6.3

Cost variable (*Y*) has been RANDOMIZED

$$\text{Pearson's } r = \frac{\Sigma z_x z_y}{N-1}$$

ID	z_{weight}	Z_{cost}	$z_x * z_y$
A	-1.34	-0.80	**POS**
B	-0.80	0.27	Neg
C	-0.27	1.34	**Pos**
D	0.27	-1.34	**Neg**
E	0.80	0.27	Pos
F	1.34	0.80	**POS**
N = 6			Σ = Pos



Magnitude

- Putting words to correlation coefficients

If r is...	Interpretation
Equal to 0	No relationship
Between 0 and 0.10	Trivial
Between 0.10 and 0.30	Small to medium
Between 0.30 and 0.50	Medium to large
Greater than 0.50	Large to very large

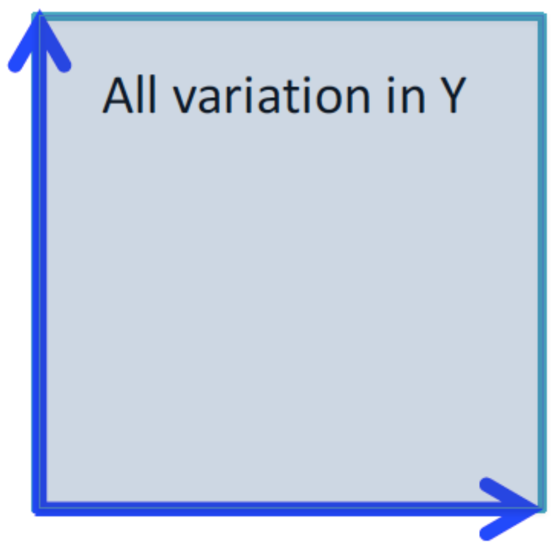
Coefficient of Determination (r^2)

- How are these two variables related? *Use r*
  - "relatedness", "covariance"

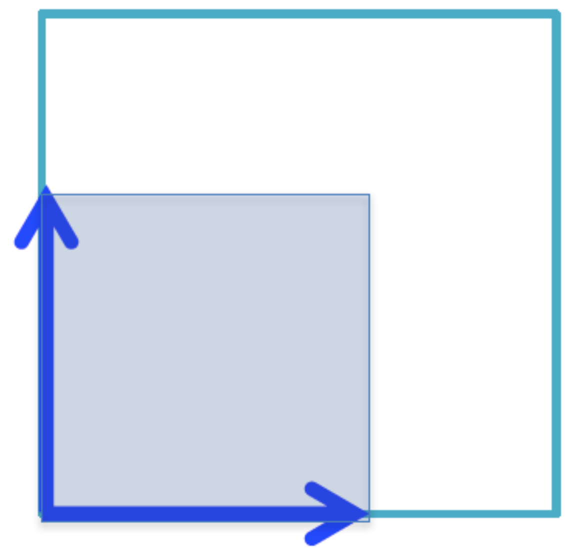
- How much variability in *Y* is accounted for by knowing *X*? *Use* r^2
  - *"explained variance"*
  - If $r = 1$, then $r^2 = 1$
    - "All (or 100%) of the variability in *Y* can be accounted for by variability in *X*"
  - When $r < 1$, then $r^2 < r$

Coefficient of Determination (r^2)

- Calculation is simple!
  - r^2 ranges from 0 to +1
    - Why not negative??



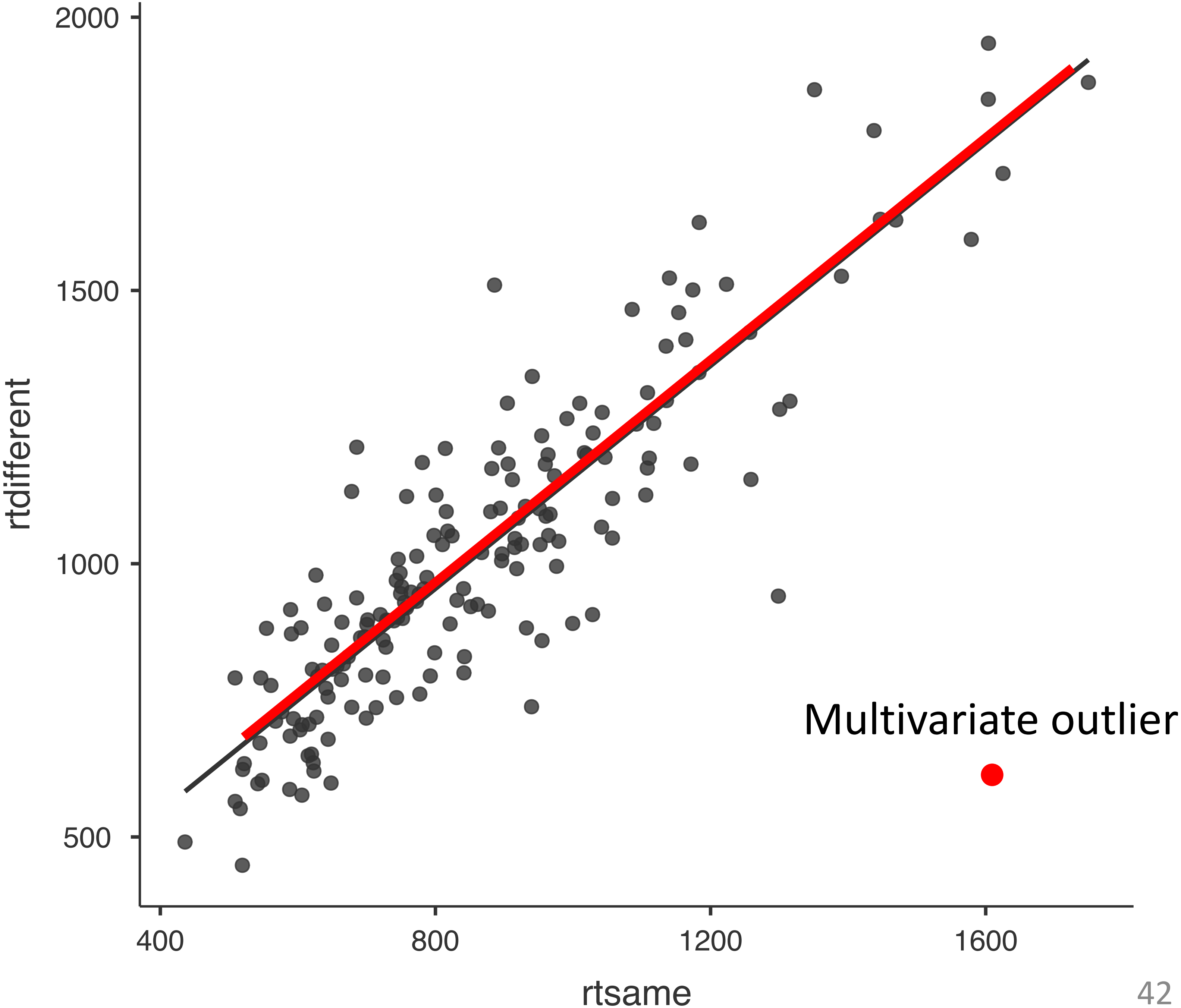
$r = 1.00$
$r^2 = 1.00$



$r = 0.65$
$r^2 = 0.42$

Other Correlation Coefficients

- **Pearson's *r*** (linear, normal dist, interval+ variable, no extreme outliers)
- **Spearman's rho** (r_s, ρ):
  - Use for ordinal data, or non-normal distributed variables
  - Only assumes variables are ranked
  - Also see Kendall's Tau (τ) & Goodman's Gamma (γ)
- **Point biserial r_b**
  - Use when 1 var is interval/ratio, 1 var is dichotomous
- **Pearson's Phi (φ)**
  - Use when both vars are dichotomous
- **Eta correlation ratio (η)**
  - Use for simple non-linear relationships
  - Common for ANOVA
  - Also see Omega correlation ratio (ω)


2000
1500
1000
500
rtdifferent
400
800
1200
1600
rtsame
Multivariate outlier