please get a handout from the back!

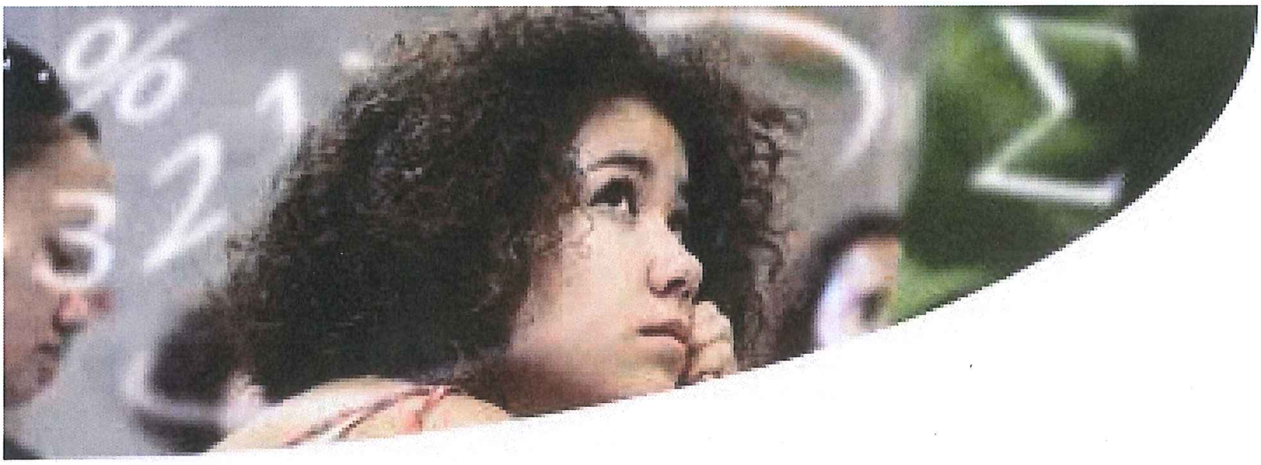
Stats 101/101G/108 Workshop

Regression and Correlation [RC]

2020

we'll start @ 10.05am...

by Leila Boyle



Stats 101/101G/108 Workshops

The Statistics Department offers workshops and one-to-one/small group assistance for Stats 101/101G/108 students wanting to improve their statistics skills and understanding of core concepts and topics.

Leila's website for Stats 101/101G/108 workshop hand-outs and information is here: **www.tinyURL.com/stats-10x**

Resources for this workshop, including pdfs of this hand-out and Leila's scanned slides showing her working for each problem are available here: **www.tinyURL.com/stats-RC**

Want to get in touch with Leila?

Leila Boyle
Undergraduate Statistics Assistance, Department of Statistics
Room 303S.288 (second floor of the Science Centre, Building 303S)
l.boyle@auckland.ac.nz; (09) 923-9045; 021 447-018

Want help with Stats?

Stats 101/101G/108 appointments

Book your preferred time with Leila here: **www.tinyurl.com/appt-stats**, or contact her directly (see above for her contact details).

Stats 101/101G/108 Workshops

Workshops are run in a relaxed environment, and allow plenty of time for questions. In fact, this is encouraged ☺

Please make sure you bring your calculator with you to all of these workshops!

- **Preparation at the beginning of the semester:**

Multiple identical sessions of a preparation workshop are run at the beginning of the semester to get students off to a good start – come along to whichever one suits your schedule!

  - Basic Maths and Calculator skills for Statistics **www.tinyURL.com/stats-BM**

- **First half of the semester**

  Five theory workshops are held during the first half of the semester:

  - Exploratory Data Analysis **www.tinyURL.com/stats-EDA**
  - Proportions and Proportional Reasoning **www.tinyURL.com/stats-PPR**
  - Observational Studies, Experiments, Polls and Surveys **www.tinyURL.com/stats-OSE**
  - Confidence Intervals: *Means* **www.tinyURL.com/stats-CIM**
  - Confidence Intervals: *Proportions* **www.tinyURL.com/stats-CIP**

- **Second half of the semester**

  Five theory workshops and one computing workshop are held during the second half of the semester:

- **Statistics Theory Workshops**

  - ✓ Hypothesis Tests: *Proportions* Ch6/7 **www.tinyURL.com/stats-HTP**
  - ✓ Hypothesis Tests: *Means (part 1)* Ch6/7 **www.tinyURL.com/stats-HTM**
  - ✓ Hypothesis Tests: *Means (part 2)* Ch8 **www.tinyURL.com/stats-HTM**
  - Chi-Square Tests Ch9 tomorrow **www.tinyURL.com/stats-CST**
  - ✓ Regression and Correlation Ch 10 today **www.tinyURL.com/stats-RC**

- **Computer Workshop:** Hypothesis Tests in *SPSS* **www.tinyURL.com/stats-HTS**

- **Useful Computer Resource:**

  If you haven't used SPSS before, try working your way through this self-paced tutorial: **www.tinyURL.com/stats-IS**

Regression and Correlation

Types of Variables

Numeric
or Scale (SPSS)
(measurements and counts)

Continuous
(measurements; few repeated values)

No gaps between possible values, if measured sufficiently accurately.

Examples

Weight; Height; Length; Temperature; Total course mark (out of 100); Age [large range]

Discrete
(usually counts; many repeated values)

Gaps between each of the values it can take.

Examples

Number of siblings; Number of cars per household; Assignment mark (out of 10); Age [small range]

Categorical
(define groups)

Nominal (SPSS)
(no idea of order)

Examples

Gender; Ethnicity; Degree (BA, BCom, BSc, etc.); Major (Chemistry, Statistics, etc.)

Ordinal (SPSS)
(fall in natural order)

Examples

Size (small, medium, large); Income bracket; Grade (A+, A, ...); Age group (<25, 25-39, 40+)

📖**Useful reference:** Chance Encounters, pages 40 – 42

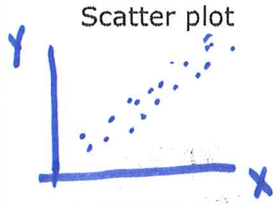
The main tool for comparing two numeric variables is the scatter plot.
What to look for in a scatter plot:

- Trend (pattern)
- Scatter
- Strength of the relationship
- Association
- Outliers
- Groupings

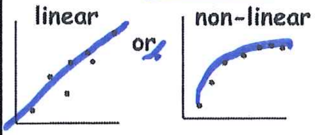
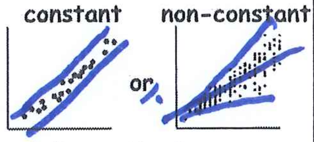
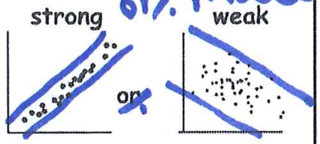
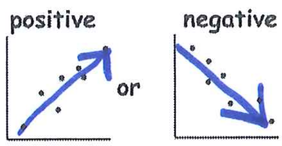
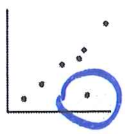
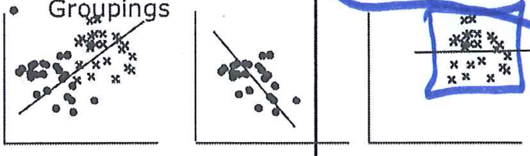
Exploring relationships between two variables

Two Numeric	Numeric versus Categorical	Two Categorical
Ch 10	Ch 6-8 (means)	Ch 9
RC	HTM (1&2)	CST/HTP

Exploratory tools to use to explore relationship/s

Two Numeric	Numeric versus Categorical	Two Categorical
Scatter plot 	Side-by-side plots on the same scale: any n { • Dot plots $n \geq 20$ { • Stem-and-leaf plots • Box plots $n \geq 50$ { • Histograms	Two-way table of counts and/or Bar graphs of proportions (on rows/columns)

none! **Features to look for:**

Two Numeric	Numeric versus Categorical	Two Categorical
• Trend 	Compare the groups by looking at:	• Most common and least common combinations
• Scatter 	• Any group differences: ◦ averages (centres) → medians, means ◦ variability (spread) → IQRs, ranges	• Differences in distributions (e.g. row and/or column bar graphs)
• Strength of relationship or/. moderate! 	◦ shapes ♦ symmetric/skewed → left/negative, right/positive ♦ modes → unimodal, bimodal, trimodal } not box!	Chi-sq test for indep.
• Association 	• Details of individual groups:	
• Outliers 	◦ outliers, gaps, clusters, groupings	
• Groupings 	Think about reasons *why* these differences, similarities and features are seen	

regression & correlation

t-test or F-test

$y = mx + c \rightarrow y = \underset{\hat{\beta}_0}{c} + \underset{\hat{\beta}_1}{m}x$

Regression

Regression looks at the relationship between two numeric variables where the two variables take on special roles:

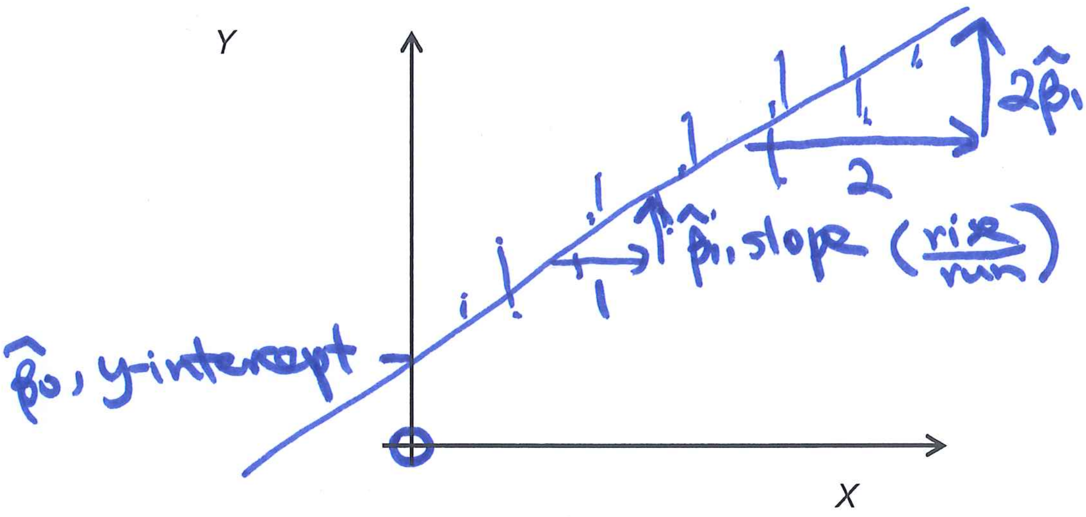
- X is used to **explain** or **predict** the behaviour of Y
- X is the **explanatory** or **independent** variable
- Y is the **dependent** or **response** variable

The two main components of the regression model are:

- **trend** and
- **scatter**.

see back page for Formulae Sheet

We use a **least squares regression line** $\boxed{\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x}$ fitted by the computer / calculator to estimate the unknown population parameters β_0 and β_1.



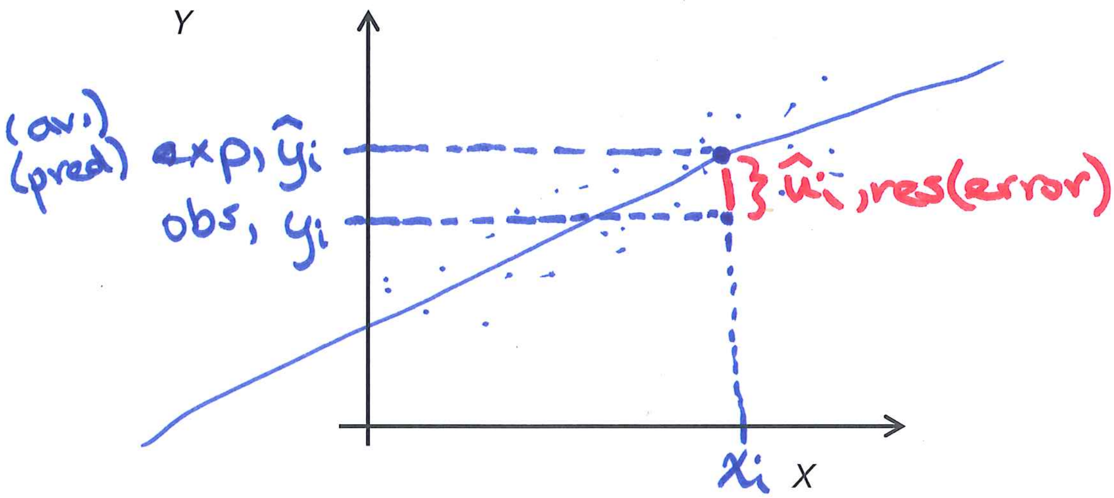
"sum", "total" (Sigma)

The single **least squares regression line** for each linear regression model:

- minimises the sum of the squared residuals/prediction errors
- has $\sum$ residuals = 0 (but so do many other lines)
- has $(\bar{x}, \bar{y})$ lying on it

- **Residuals**
  - Errors, residuals or prediction errors are all terms for the same thing.
  - A residual is the (vertical) distance between the **actual observed value** y_i and the **expected estimated value** $\hat{y}_i$, i.e.:

Errors = observed – expected ($\hat{u}_i = y_i - \hat{y}_i$) memorise!



- **Hypotheses**

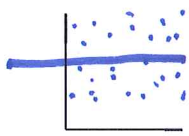
Recall the *t*-test?

see back page for Formulae Sheet

Use: $\boxed{t_0 = \dfrac{\text{estimate} - \text{hypothesised value}}{\text{std error}}}$

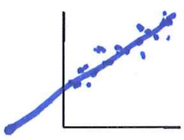
where in this case the degrees of freedom are: $\boxed{df = n - 2}$

H_0: $\beta_1 = 0$ (there **is no** **linear relationship**)

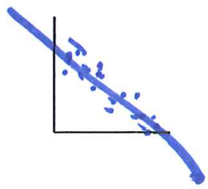


Slope

H_1: $\beta_1 \neq 0$ (there **is a** **linear relationship**)



or,



In SPSS, the output always comes out in the same way:

predicted, average, expected

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

$df = n-2$

Regression

Coefficients(a)

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
1 (Constant) β_0	$\hat{\beta}_0$	$se(\hat{\beta}_0)$		t_0	*p-value*
***X*-axis_Variable**	$\hat{\beta}_1$	$se(\hat{\beta}_1)$	*r*	t_0	*p-value*

a Dependent Variable: ***Y*-axis_Variable**

y-intercept

slope, β_1

$H_0: \beta_1 = 0$

vs $H_1: \beta_1 \neq 0$

$\rightarrow t_0 = \dfrac{\text{est} - \text{hyp val}}{\text{std err}} = \dfrac{\hat{\beta}_1 - 0}{se(\hat{\beta}_1)}$

Correlations

		***X*-axis_Variable**	***Y*-axis_Variable**
***X*-axis_Variable**	Pearson Correlation	1	*r*
	Sig. (2-tailed)		*p-value*
	N	*n*	*n*
***Y*-axis_Variable**	Pearson Correlation	*r*	1
	Sig. (2-tailed)	*p-value*	
	N	*n*	*n*

Recall: The P-value:

- In regression, we are carrying out *t*-tests, just like in Chapter 7 and 8!
- Therefore, the ***P*-value:**
  - ✓ is the **conditional** probability of observing a test statistic as extreme as that observed or more so, **given that** the null hypothesis, H_0, is true.
  - ✓ is the probability that sampling variation would produce an estimate that is at least as far from the hypothesised value than the estimate we obtained from our data, **assuming that** the null hypothesis, H_0, is true.
  - ✓ measures the strength of evidence **against** H_0.

saying the same thing!

memorise!

- We interpret the *P-value* as a **description** of the **strength of evidence against the null hypothesis, H_0**. The **smaller** the *P-value*, the **stronger** the evidence against H_0:

P-value	Evidence against H_0
> 0.10	None
≈ 0.07	Weak
≈ 0.05	Some
≈ 0.01	Strong
≤ 0.001	Very Strong

memorise!

- An alternative approach often found in research articles and news items is to describe the test result as (statistically) significant or not significant. A test result is said to be significant when the *P-value* is "small enough"; usually people say a *P-value* is "small enough" if it is less than 0.05 (5%):

Testing at a 5% level of significance:

P-value	Test result	Action
< 0.05	Significant	Reject H_0 in favour of H_1
> 0.05	Nonsignificant	Do not reject H_0

✓ .049
✗ .051

Testing can be done at any level of significance; 1% is common but 5% is what most researchers use.

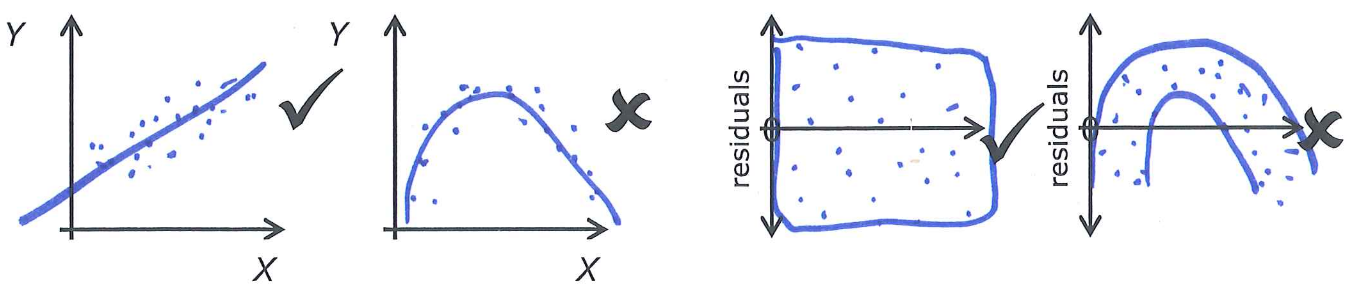
The level of significance can be thought of as a false alarm error rate, i.e. it is the proportion of times that the null hypothesis will be rejected when it is actually true (which can result in action being taken when really no action should be taken).

Thus, a statistically significant result means that a study has produced a "small" *P-value* (usually < 5%).

- **Assumptions** of simple linear regression are:
  1. There is a **linear** relationship between X and Y.
  2. Errors are all **independent**. can't be checked!
  3. Errors are **Normal**ly distributed (with $\mu = 0$).
  4. Errors all have the **same std deviation**, σ, regardless of the value of x.
- **Assumption checking using plots of the data and residual plots**

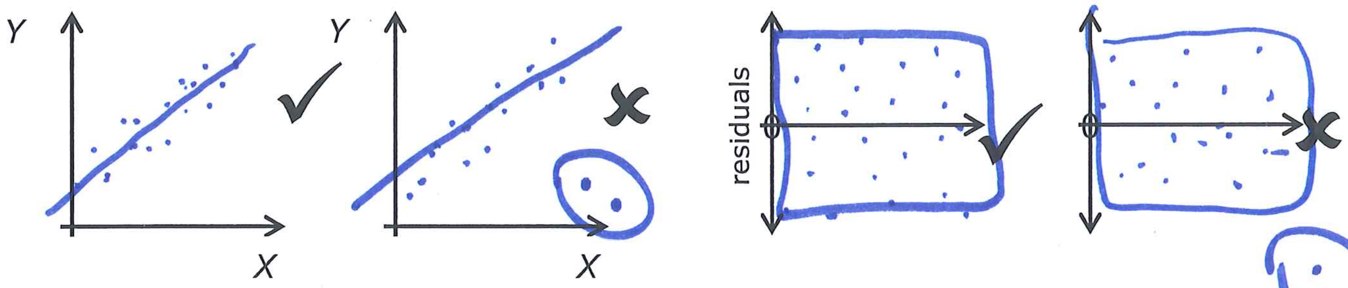
1. There is a **linear** relationship between X and Y.

Scatterplot of data: Residual plot:



3. Errors are **Normal**ly distributed (with $\mu = 0$).

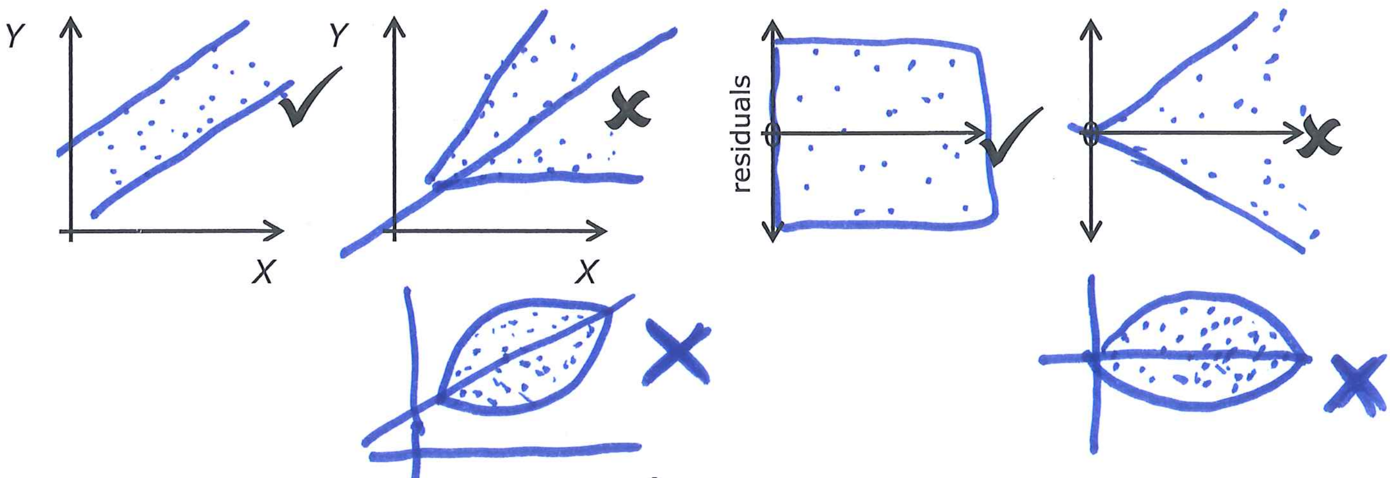
Scatterplot of data: Residual plot:



constant scatter | spread | variation

4. Errors all have the **same std deviation**, σ, regardless of the value of x.

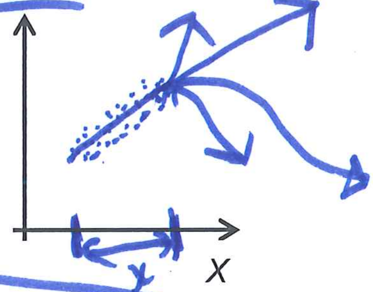
Scatterplot of data: Residual plot:



11.20-11.35am – have a break! restart on pg 10 @ 11.35...

- **Estimating / Predicting**

✓ Within the range of our observed X-values this can be done with confidence. Predicting outside the range of our observed X-values is dangerous. A relationship that fits the data well may not extend outside that range.



✓ **Confidence Interval (for the mean)**

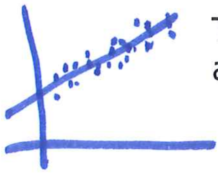
This estimates the **mean** Y-value at a specified value of x.

The width of the interval allows for:

CI

- uncertainty about the values of β_0 and β_1.

$\boxed{\text{estimate} \pm t \times se(\text{estimate})}$



✓ **Prediction Interval**

This predicts the Y-value for **an individual** with a specified value of x.

PI

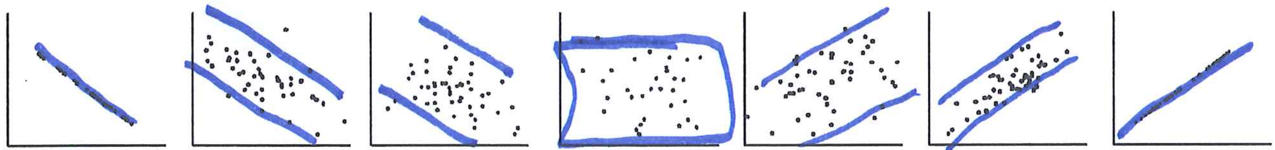
The width of the interval allows for:

- uncertainty about the values of β_0 and β_1 **and**
- uncertainty due to the random scatter about the line.

✓ For a given value of x, the **95% prediction interval** (PI) is **always wider** than the **95% confidence interval for the mean** (CI).

✓ **The Sample Correlation Coefficient, *r***

✓ ***r*** has a value between -1 and +1:



$r = -1$ $\quad$ $r = -0.7$ $\quad$ $r = -0.4$ $\quad$ $r = 0$ $\quad$ $r = 0.3$ $\quad$ $r = 0.8$ $\quad$ $r = 1$

- $r = -1$, then X and Y have a **perfect negative linear relationship**
- $r = 0$, then X and Y have **no linear** relationship but they may have **some other non-linear relationship**
- $r = 1$, then X and Y have a **perfect positive linear relationship**

✓ ***r*** measures the **strength** and **direction** of the **linear** association between **two numeric variables**

✓ ***r*** measures how close the points **come to lying on a straight line**

✓ The value of ***r*** is the same if the axes are swapped around – it doesn't matter which variable is X and which one is Y

✓ ***r*** has no units → a computer / calculator can give you the value of ***r***

✓ **Correlation DOES NOT imply causation**