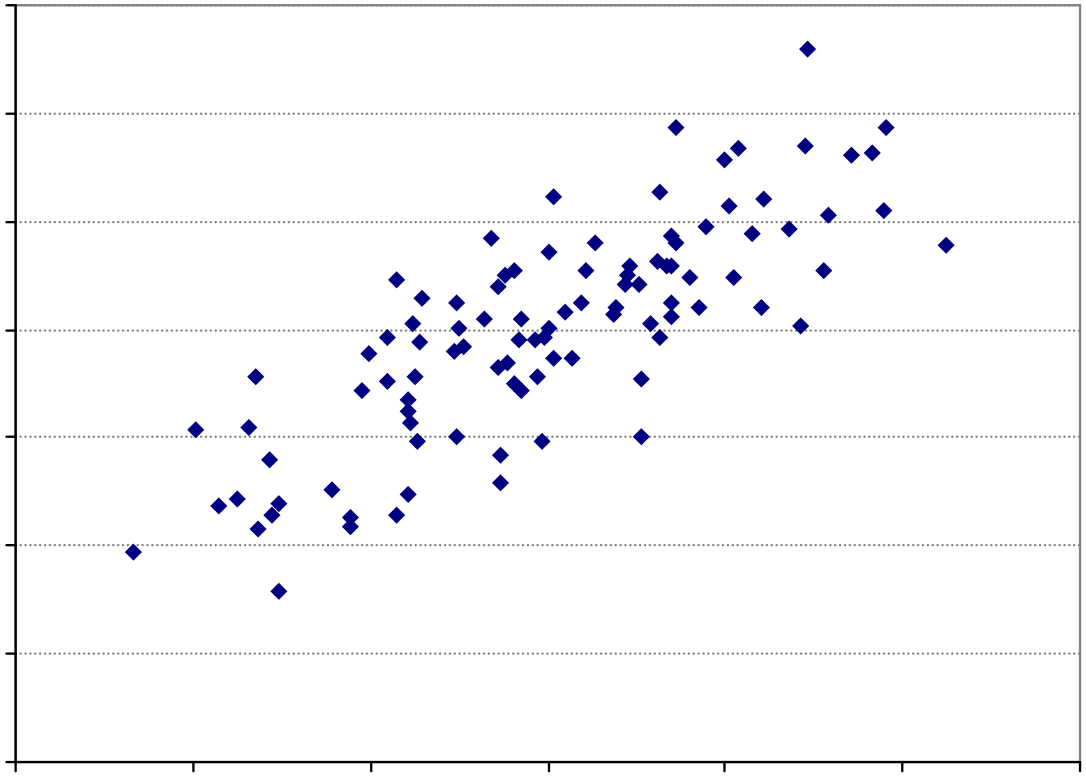
Chapter 7: Correlation and regression

- Correlation and regression techniques examine the relationships between variables, e.g. between the price of doughnuts and the demand for them.

- Such analyses can be useful for formulating policies, e.g. if there is a positive relationship between the quantity of money and the price level, inflation may be curbed by stricter control of the money supply.

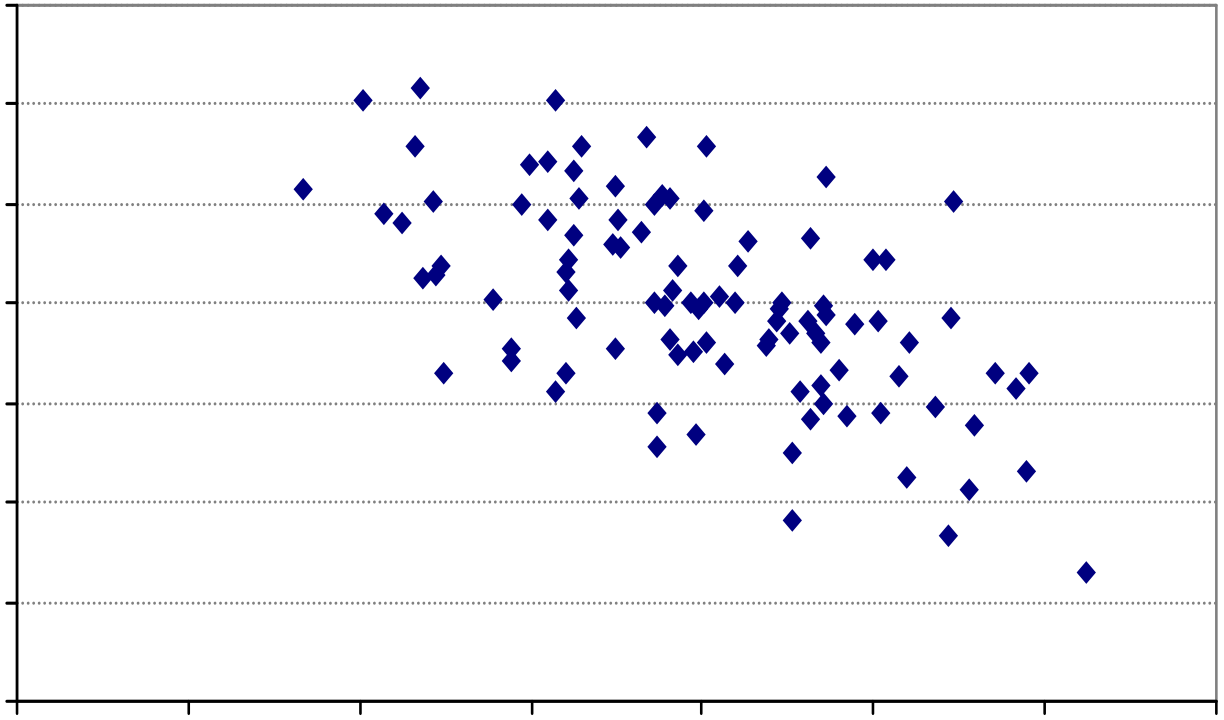
Slide 7.2

Positive correlation



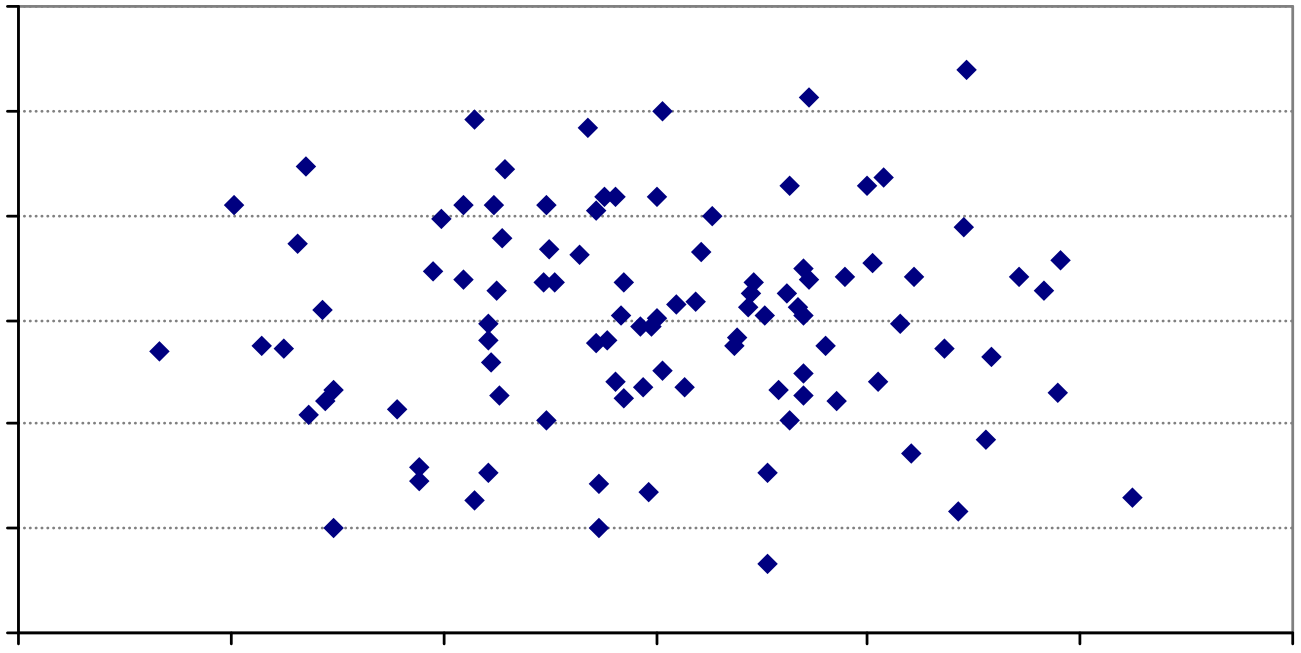
E.g. income and food expenditure

Negative correlation



E.g. demand and price

Zero (absence of) correlation



E.g. sales of cameras and the price of fish

The correlation coefficient

- Measures the strength of association between two variables, *X* and *Y*.
- $-1 \leq r \leq +1$
- Positive correlation: $r > 0$
- Negative correlation: $r < 0$
- Zero correlation: $r \approx 0$

The correlation coefficient (continued)

- The closer *r* is to +1 (or -1), the closer the points lie to a straight line with positive (negative) slope.
- Slide 7.2: $r = 0.8$
- Slide 7.3: $r = -0.7$
- Slide 7.4: $r = 0$

Formula for the correlation coefficient

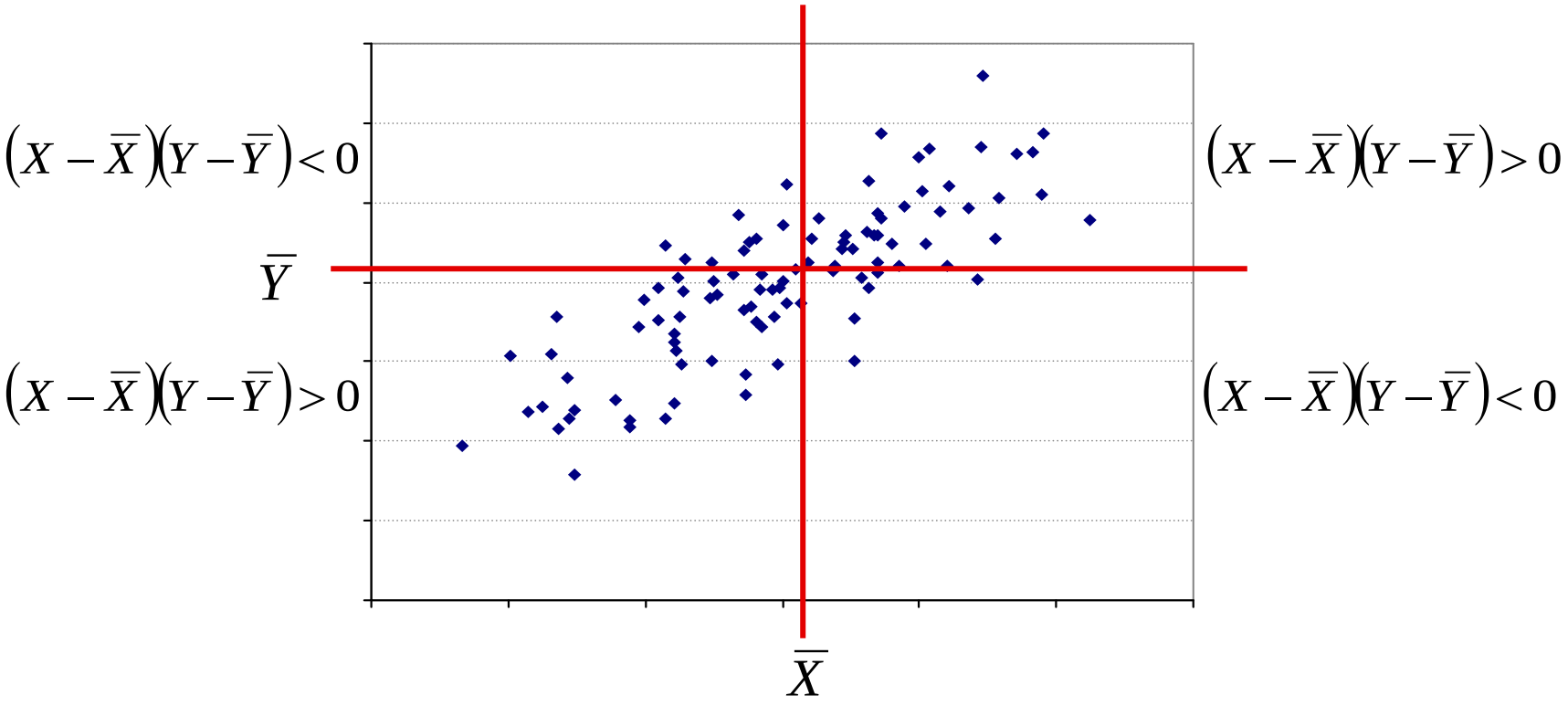
- Use either

$$r = \frac{\sum \left(X - \bar{X}\right)\left(Y - \bar{Y}\right)}{\sqrt{\sum \left(X - \bar{X}\right)^2 \times \sum \left(Y - \bar{Y}\right)^2}}$$

- or, equivalently

$$r = \frac{n\sum XY - \sum X \sum Y}{\sqrt{(n\sum X^2 - (\sum X)^2)(n\sum Y^2 - (\sum Y)^2)}}$$

Why the formula works



More >0 than <0 points, hence $r > 0$

Slide 7.9

Calculation of *r* between growth and birth rates

Country	Birth rate	GNP growth			
	Y	X	Y^2	X^2	XY
Brazil	30	5.1	900	26.01	153.0
Colombia	29	3.2	841	10.24	92.8
Costa Rica	30	3.0	900	9.00	90.0
India	35	1.4	1,225	1.96	49.0
Mexico	36	3.8	1,296	14.44	136.8
Peru	36	1.0	1,296	1.00	36.0
Philippines	34	2.8	1,156	7.84	95.2
Senegal	48	–0.3	2,304	0.09	–14.4
South Korea	24	6.9	576	47.61	165.6
Sri Lanka	27	2.5	729	6.25	67.5
Taiwan	21	6.2	441	38.44	130.2
Thailand	30	4.6	900	21.16	138.0
Total	380	40.2	12,564	184.04	1,139.7

Calculation of *r* between growth and birth rates (continued)

- Using the second formula,

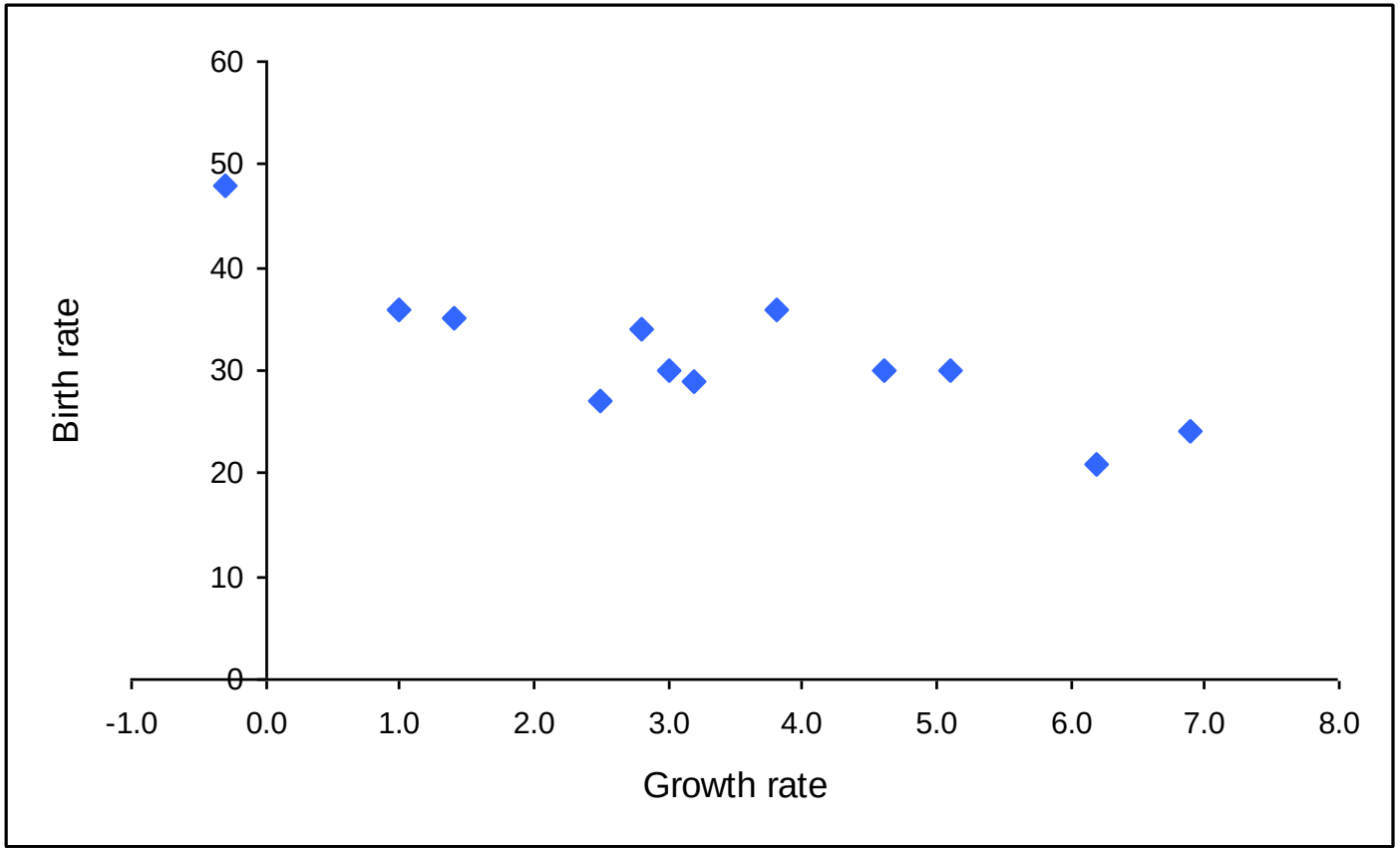
$$r = \frac{n\sum XY - \sum X \sum Y}{\sqrt{(n\sum X^2 - (\sum X)^2)(n\sum Y^2 - (\sum Y)^2)}}$$

- we obtain

$$r = \frac{12 \times 1{,}139.7 - 40.2 \times 380}{\sqrt{(12 \times 184.04 - 40.2^2)(12 \times 12{,}564 - 380^2)}} = -0.824$$

Slide 7.11

Chart of birth rate against growth rate



Notes about *r*

- The correlation between *Y* and *X* is the same as between *X* and *Y*, i.e. it does not matter which variable is labelled *X* and which *Y*.

- *r* is independent of units of measurement. If the birth rate were measured as births per 100 population (3.0, 2.9,...) *r* would still be -0.824.

- Correlation does not imply causality.

Is the result statistically significant?

- H_0: $\rho = 0$
  H_1: $\rho \neq 0$

- The null asserts no genuine association between *X* and *Y* and the sample correlation observed is just due to (bad) luck.

- The test statistic is $t = \dfrac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t_{n-2}$

Testing the hypothesis

- Choose α = 5%, hence t^*_{10} = 2.228

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{-0.824\sqrt{12-2}}{\sqrt{1-(-0.824)^2}} = -4.59$$

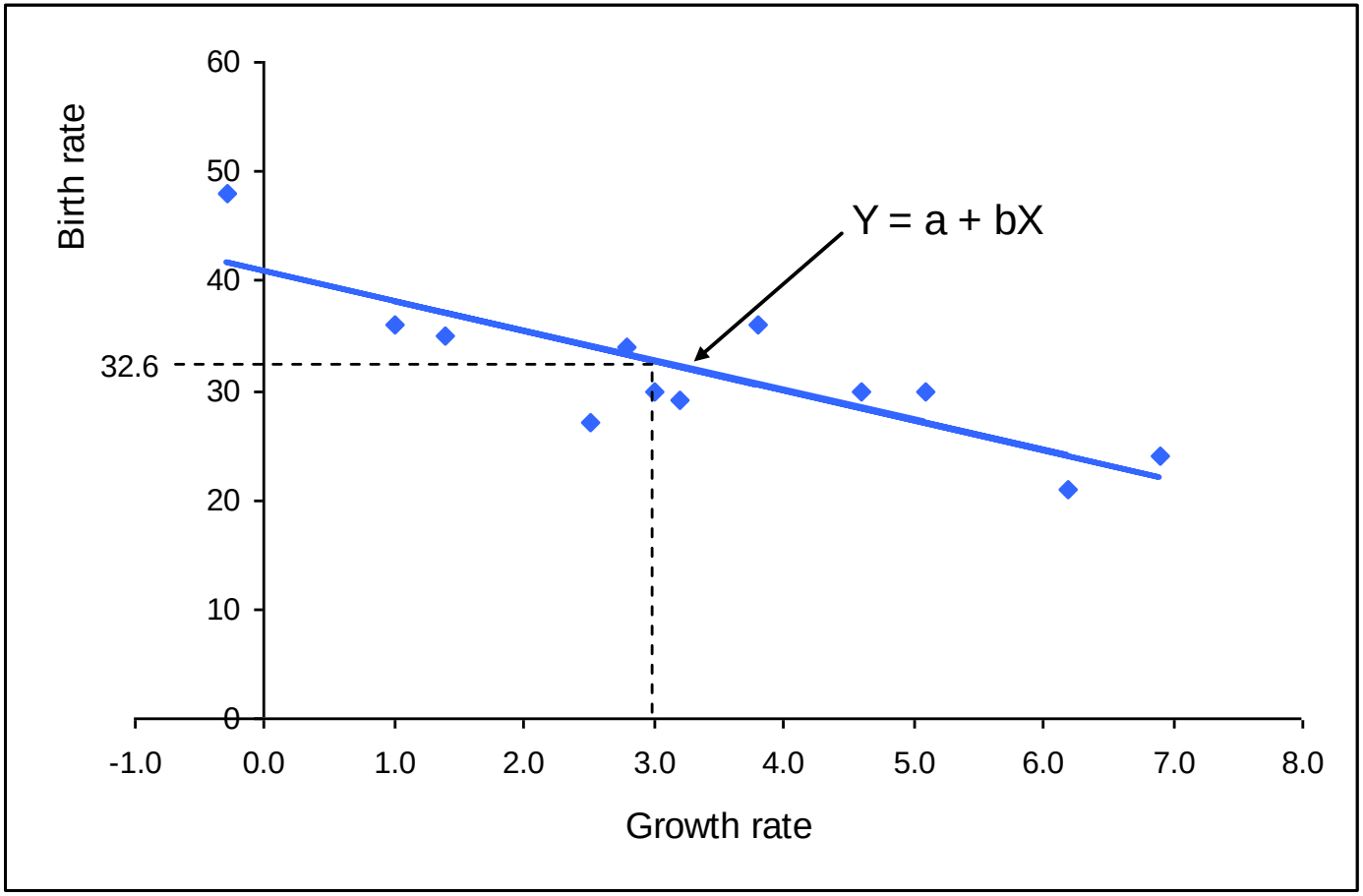
- Hence we reject H_0. There does seem to be some genuine association between *X* and *Y*.

Regression

- We now assert *X* causes *Y*, i.e. the growth rate influences the birth rate (not vice versa)
- Regression measures the effect of *X* upon *Y* and whether it is statistically significant
- Regression also allows several explanatory variables to influence *Y*

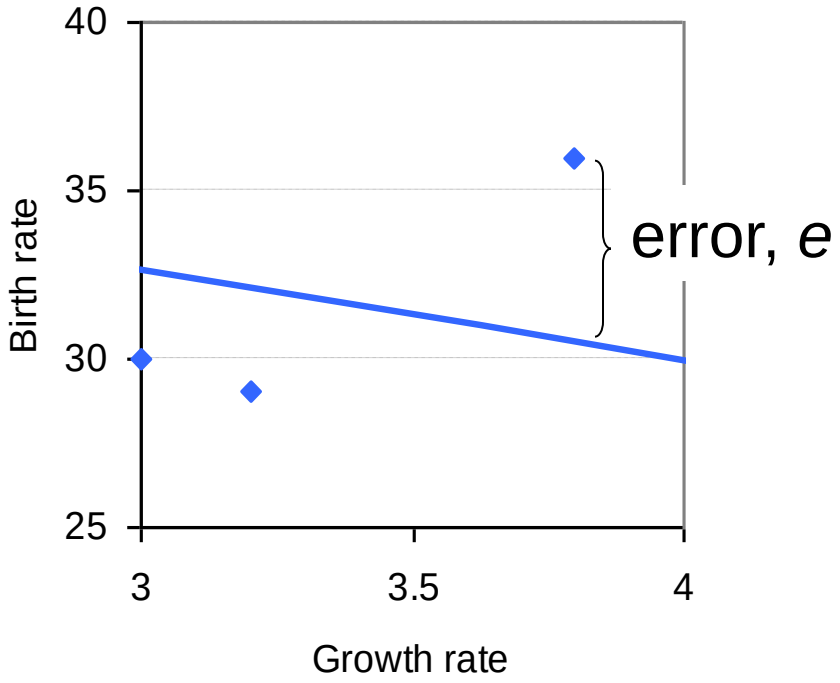
Slide 7.16

The regression line



How to obtain the regression line

- Minimise the sum of squared errors, Σe^2
- The error is the vertical distance between an observation and the regression line



Regression formulae

- Slope

$$b = \frac{n\sum XY - \sum X \sum Y}{n\sum X^2 - (\sum X)^2}$$

- Intercept

$$a = \bar{Y} - b\bar{X}$$

Calculation of regression coefficients

- Slope

$$b = \frac{12 \times 1{,}139.70 - 40.2 \times 380}{12 \times 184.04 - 40.2^2} = -2.700$$

- Intercept

$$a = \frac{380}{12} - (-2.700) \times \frac{40.2}{12} = 40.711$$

- $Y_i = 40.71 - 2.70X_i + e_i$

Measuring goodness of fit

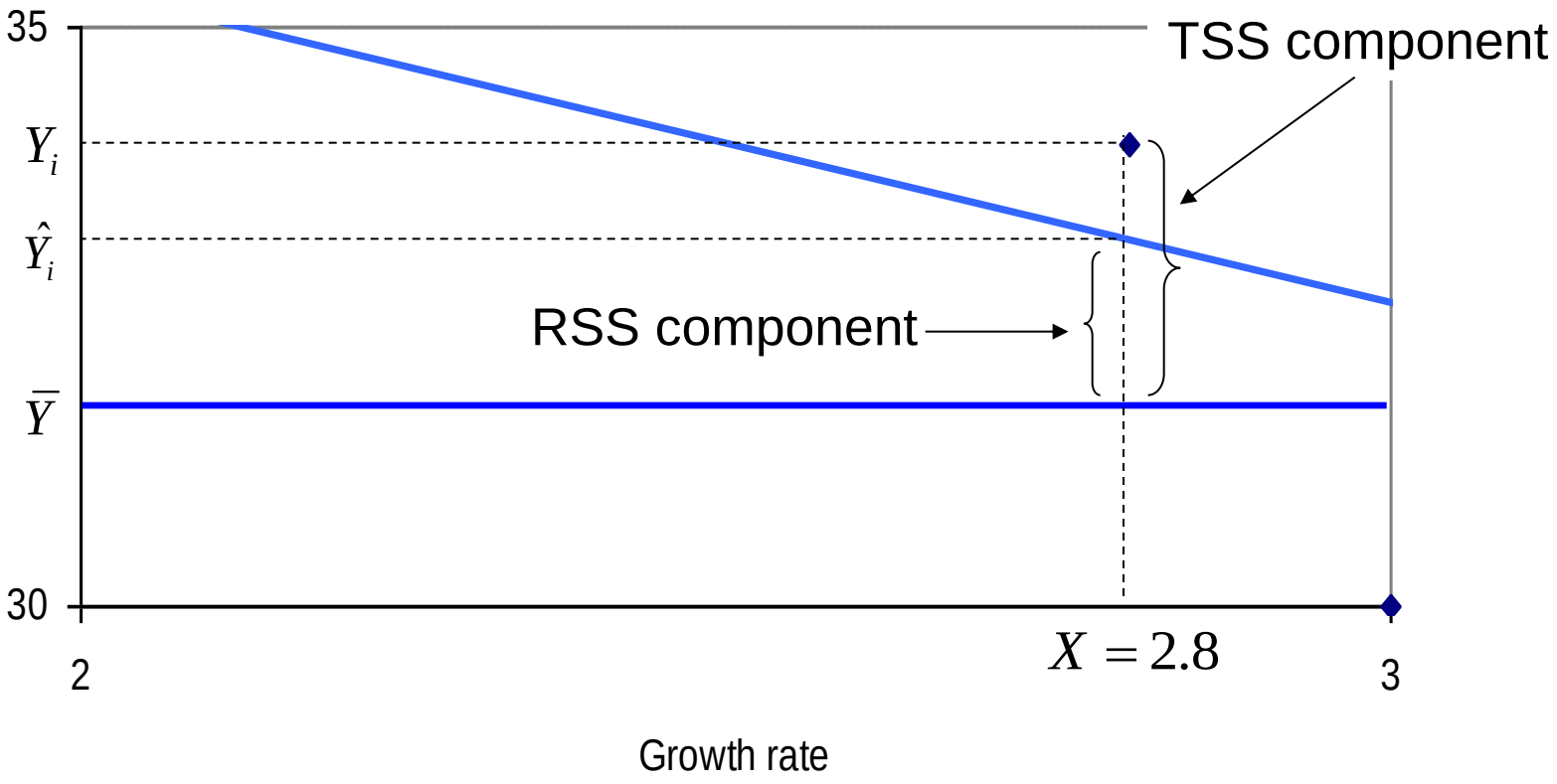
- Use the coefficient of determination, R^2

$$R^2 = \frac{RSS}{TSS}$$

- $0 \leq R^2 \leq 1$

- RSS: regression sum of squares
  TSS: total sum of squares

The component parts of R^2



Calculating sums of squares

- TSS = RSS + ESS

$$TSS = \sum \left(Y - \bar{Y}\right)^2 = \sum Y^2 - n\bar{Y}^2 = 12,564 - 12 \times 31.67^2 = 530.67$$

$$ESS = \sum \left(Y - \hat{Y}\right)^2 = \sum Y^2 - a\sum Y - b\sum XY$$

$$= 12,564 - 40.71 \times 380 - \left(-2.7\right) \times 1,139.7 = 170.75$$

$$RSS = 530.67 - 170.75 = 359.92$$

- Hence $R^2 = \dfrac{359.92}{530.67} = 0.678$

Summary

- Correlation measures the association between two variables
- Regression extends this by measuring the effect of *X* upon *Y* (the slope coefficient *b*)
- The regression line is found by minimising the sum of squared errors. It is the 'line of best fit'
- A measure of goodness of fit is R^2