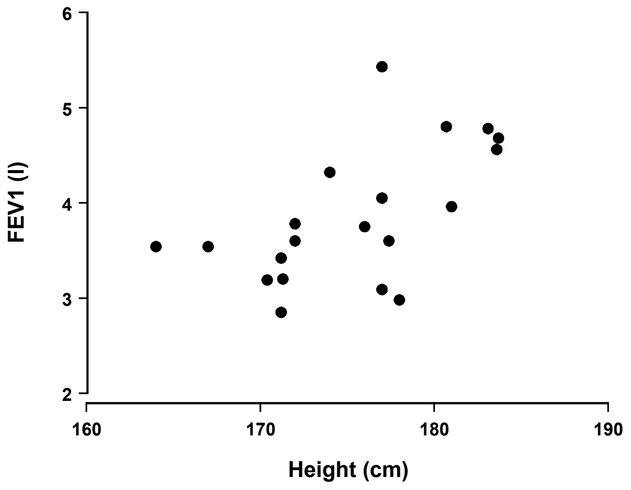
Relationship between Two Continuous Variables

- In all applications, start with scatterplot

- Example: 20 male students, heights (cm) and FEV1 (l) (Bland)



Questions that may arise:

- Is there a relationship between Height and FEV1?

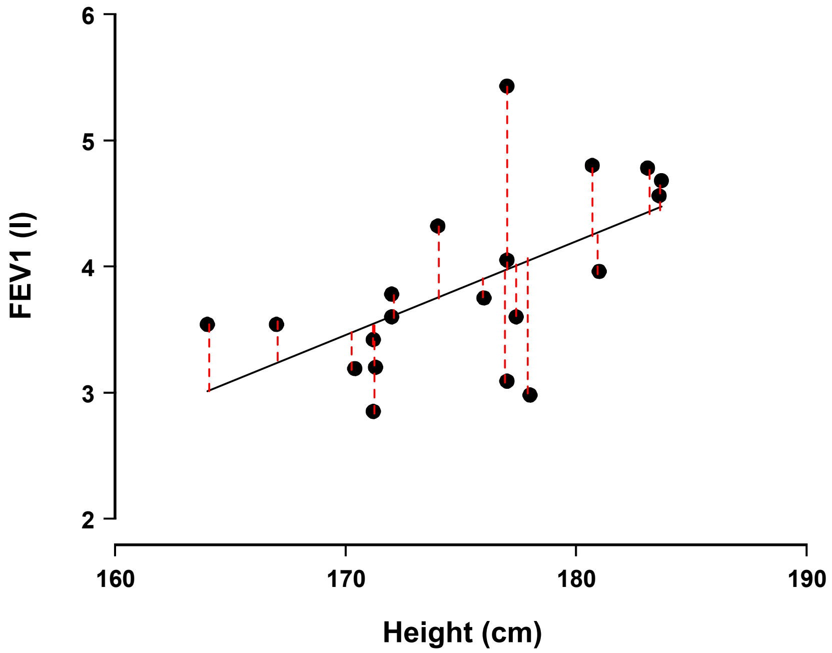
- How should it be quantified?

- Can FEV1 be predicted from height?

- If so, how well?

Usual Approach

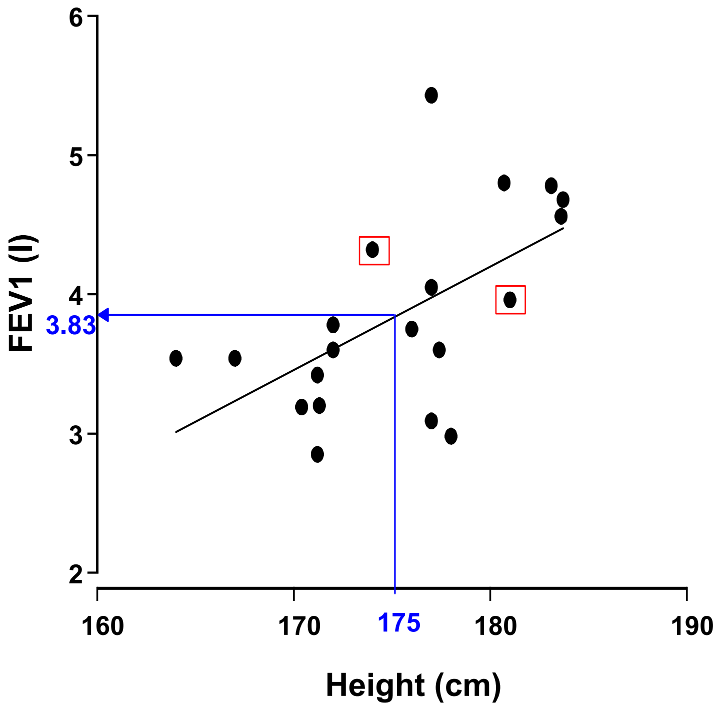
Find “line of best fit”, or “least squares fit”



Line minimises sum of squares of red lengths

But what does the line mean?

Interpretation of line of best fit



Use line for prediction at new heights:
e.g. student of height 175cm is predicted to have FEV1 of 3.83 l

But some students are taller but have lower FEV1s (e.g. boxed points)

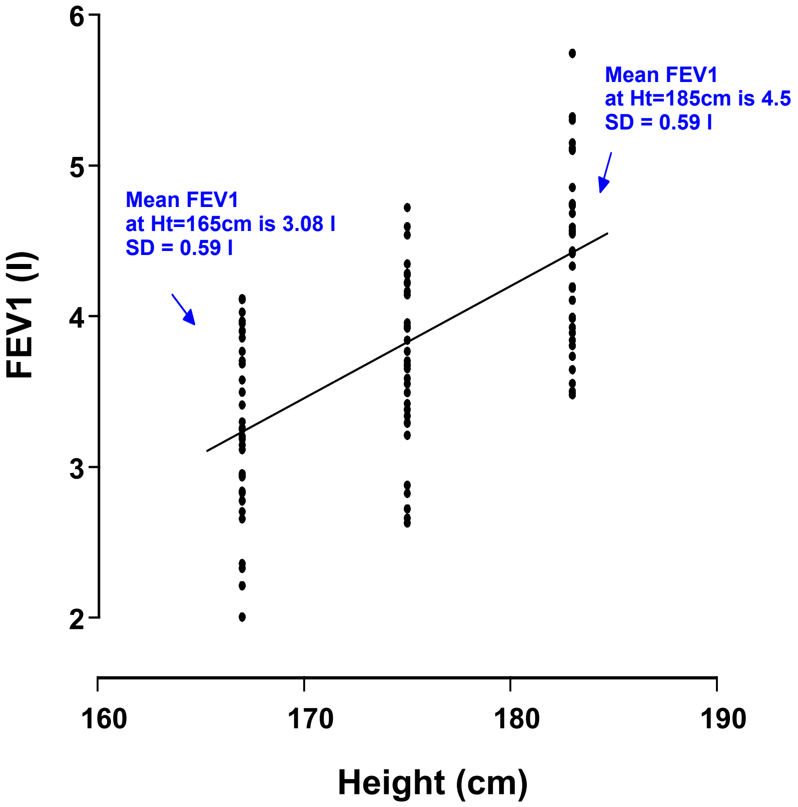
So what does the line mean?

Statistical Model I

- So far, fitting the line has been a geometrical exercise
- What is the statistical basis of the exercise?
- Recall case of one variable:

  values are distributed about an unknown mean, μ, with unknown SD σ; these are the *population parameters*. We *estimate* μ and σ by sample mean, *m* and SD, *s*.
- FEV1 is assumed to vary about a mean $\mu(Height)$ *which depends on height*, and has SD σ

Statistical Model II



- Dependence of mean of FEV1 on height is shown on left for some artificial data

- Dependence of mean of FEV1 on height known as the *Regression of FEV1 on Height*

- Technique usually known as *simple linear regression* as more complicated variants are possible

What are the Parameters?

In simple linear regression, mean of FEV1 (y-variable) is related to Height (x variable) by a straight line.

- This has mathematical expression

  Mean FEV1 = $\alpha + \beta \times$ Height

  Known as the *Regression Line*: parameters are α & β

- Also, there is σ, the SD measuring how much FEV1s vary about the regression line.

Quantities Estimated in Simple Linear Regression

Parameter	Meaning	Estimate	Units
α	Intercept	a	litres (y)
β	Slope rate of change y with x	b	litres/cm (y per x)
σ	SD about line	s	litres (y)

Also, the standard error of b (i.e. how precise b is as an estimate of β) is also important.

Is there a relationship at all?

Relationship is:

Mean FEV1 = $\alpha + \beta \times$ Height

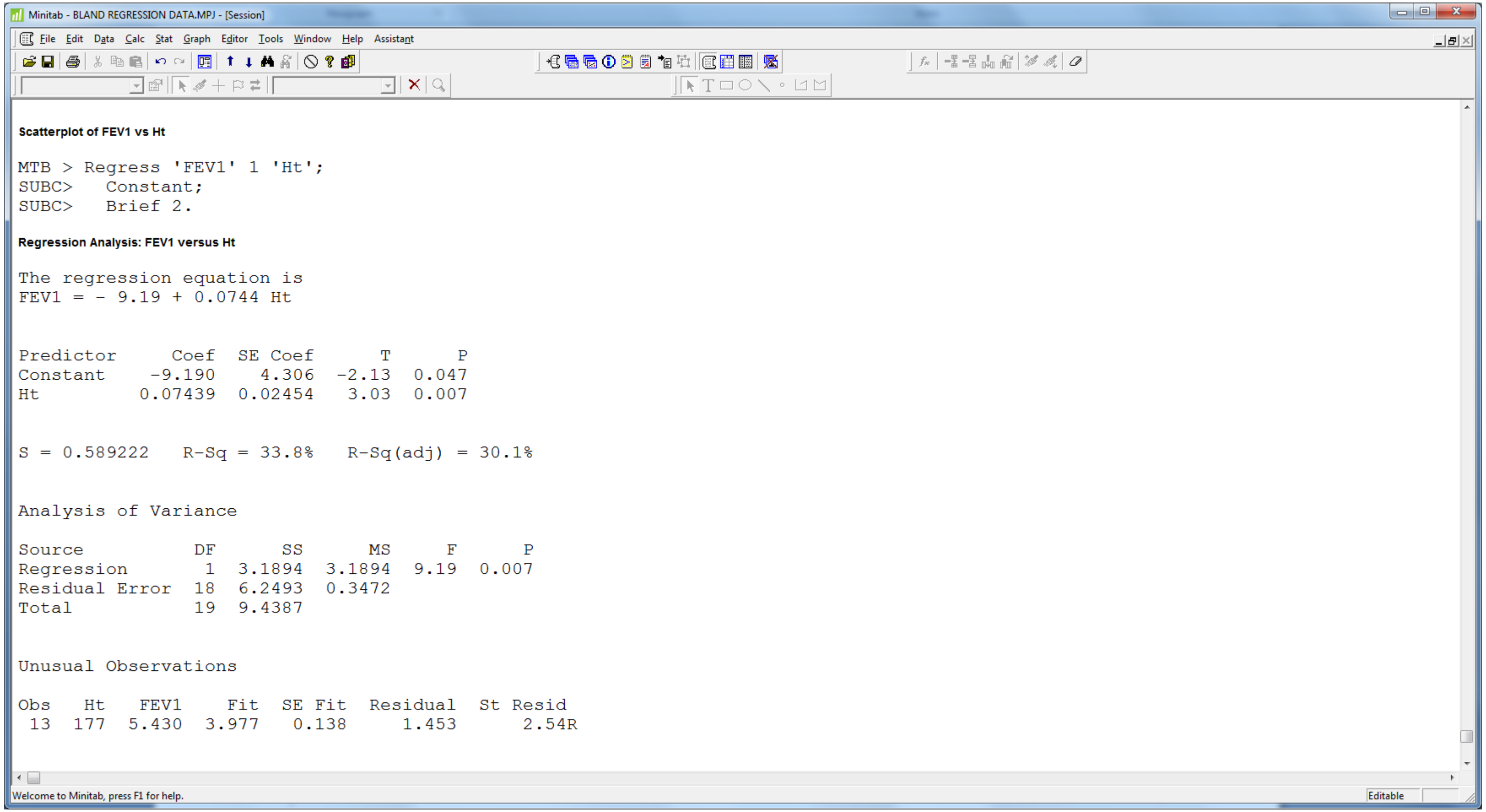
In particular, if $\beta = 0$ then mean FEV1 is unaffected by Height, i.e. there is no straight-line relationship.

Of course, β is unknown, we only know b.

Is the value of b compatible with $\beta = 0$?

Test null hypothesis $\beta = 0$ by referring $\dfrac{\text{b}}{\text{se of b}}$ to *t* distribution (n-2) degrees of freedom and obtain P-value

Minitab Analysis of FEV1 and Height


Minitab - BLAND REGRESSION DATA.MPJ - [Session]
File Edit Data Calc Stat Graph Editor Tools Window Help Assistant
Welcome to Minitab, press F1 for help.
Editable

Scatterplot of FEV1 vs Ht

```
MTB > Regress 'FEV1' 1 'Ht';
SUBC>   Constant;
SUBC>   Brief 2.
```

Regression Analysis: FEV1 versus Ht

```
The regression equation is
FEV1 = - 9.19 + 0.0744 Ht


Predictor     Coef  SE Coef      T      P
Constant    -9.190    4.306  -2.13  0.047
Ht         0.07439  0.02454   3.03  0.007


S = 0.589222   R-Sq = 33.8%   R-Sq(adj) = 30.1%


Analysis of Variance

Source          DF      SS      MS     F      P
Regression       1  3.1894  3.1894  9.19  0.007
Residual Error  18  6.2493  0.3472
Total           19  9.4387


Unusual Observations

Obs   Ht   FEV1    Fit  SE Fit  Residual  St Resid
 13  177  5.430  3.977   0.138     1.453     2.54R
```

FEV1 is response, height is ‘predictor’: a=-9.19 l, b=0.074 l/cm(se = 0.025 l/cm), s=0.59 l
b/(se of b) = 3.03 giving P=0.007, so there is good evidence of a relationship between FEV1 & Height.

Interpretation of Output

- Intercept a=-9.19 l: FEV1 at zero height. Needs to be in equation but is usually of little direct interest (in any case relationship will not hold right down to Height = 0). In particular, Minitab's P value for a (labelled 'Constant') is a test of a hypothesis that is of no interest.

- Slope b = 0.074 l/cm: this is main parameter of interest and quantifies relationship: mean FEV1 increases by 0.074 l for each cm increase in height.

- SD about line s = 0.59 l: individuals of a fixed height all share same mean and this quantity describes the spread of individual FEV1 values about that mean.

- Test of $\beta = 0$: gives P =0.007, indicating observed relationship is unlikely to be due to chance.

Prediction

Given a male student of height *h* (e.g. 180 cm), what is his predicted FEV1?

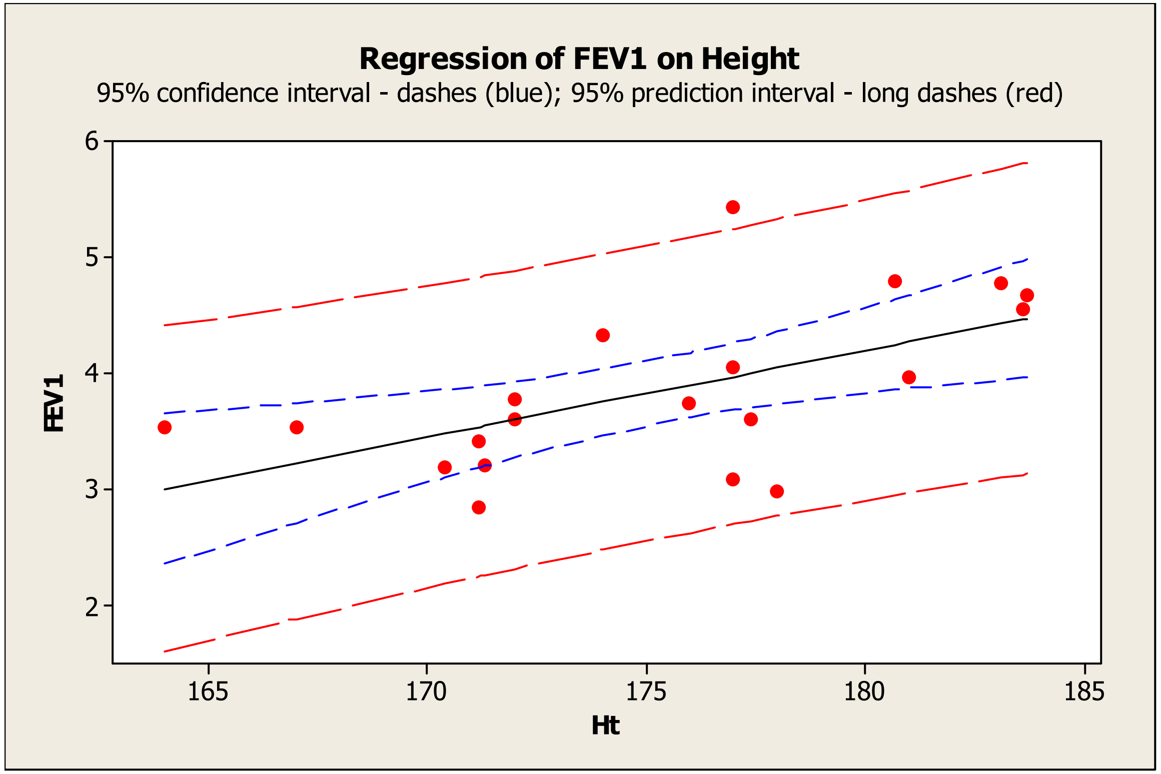
Use the mean FEV1 of students of this height as prediction namely:

$a + b \times h$, i.e. in this example $-9.19 + 0.07439 \times 180 = 4.20$l

But what if we wanted to know likely upper and lower limits?

In regression there are two related limits and care is needed to understand difference and choose the correct one.

Prediction Limits



- Narrower (more “curved”) limits give 95% confidence interval for <u>mean</u> at given height
- Wider limits are 95% prediction intervals for <u>individuals</u> of given height (e.g. 180cm)
- Former, but not latter, will get narrower as sample size increases
- Confidence and prediction intervals can be found from Minitab (Options button in regression dialog box), giving prediction interval (2.91, 5.49) (CI is 3.83, 4.57)

Assumptions

1. Mean y is linearly related to x: i.e. Mean FEV1 = α + β Height

2. Variation about this mean has <u>constant</u> SD, σ

3. For significance test and confidence/prediction intervals, the variation about the line is assumed to follow a Normal Distribution

Note: no assumption of Normality, randomness etc needed for the x variable.

Checking Assumptions

1. (Mean is linearly related to x variable). Not as easy as might be thought, because of random variation about the line. Essential to plot the data and try to decide if anything other than a straight line is needed.

2. (Constant SD about the line). Calculate the residuals, one for each point on the graph. The residual is the vertical distance of point from line (points above line +ve, below -ve). Minitab can be asked to save these. Plot residuals versus x variable and see if spread looks constant.

3. (Normality). It is actually the residuals that are assumed to be Normal, so these must be tested for Normality. Use Probability plot (Under **Graph** menu on Minitab)

Pitfalls

- Relationship is empirically based, i.e. you are using evidence from the sample to quantify a relationship between x and y.

- So do not apply relationship to quantities that could not have come from your sample.

E.g. do not use for values of x,y outside the ranges in your sample. Do not use FEV1 vs height line for children, females, older males.

- Regressing FEV1 on Height gives "FEV1 = a + b Height". Do not expect Regression of Height on FEV1 to give "Height = -a/b +FEV1/b", because it does not! Need to decide which variable is y and which is x; this is not always easy and depends on context

- Beware of outliers: these are points which do not seem to fit and can distort fitted line very badly. Need to decide if point is genuine. If it is, you are probably in trouble.

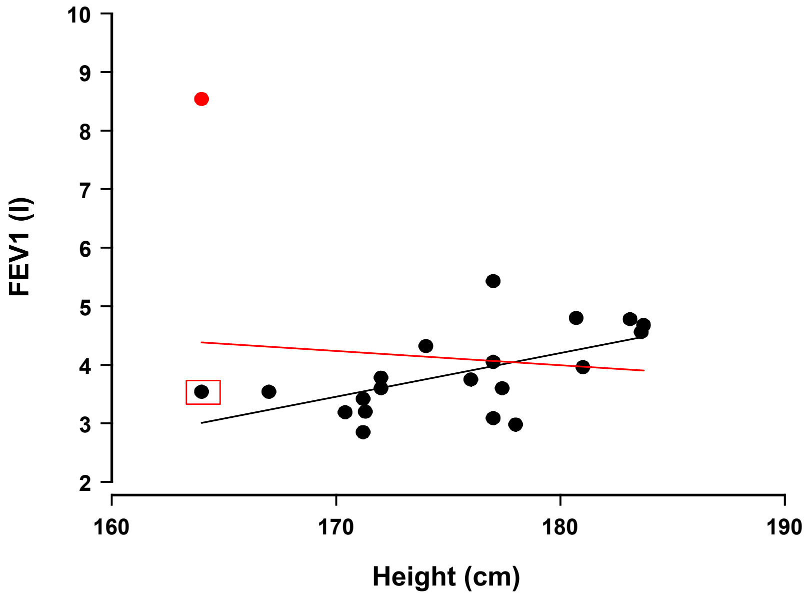


Figure shows effect of a transcription error:

FEV1 of boxed point is 3.54 l

Point with large FEV1 has been entered as 8.54 l

Using erroneous value line now slopes downwards!

Correlation

So far the ‘strength’ of the relationship between the variables has not been considered directly.

- This is what the *Correlation Coefficient* measures
- Sample correlation is usually written as *r*
- Sometimes called Pearson’s r, or product-moment correlation coefficient
- Applicable to pairs of continuous variables

Values of r

- r can only take values between -1 and +1

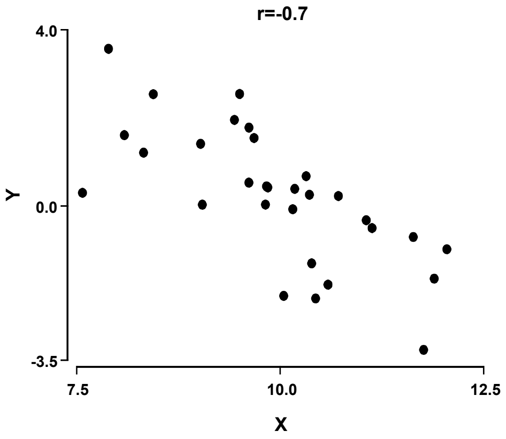
- Value of +1 means points lie exactly on a straight line with positive slope

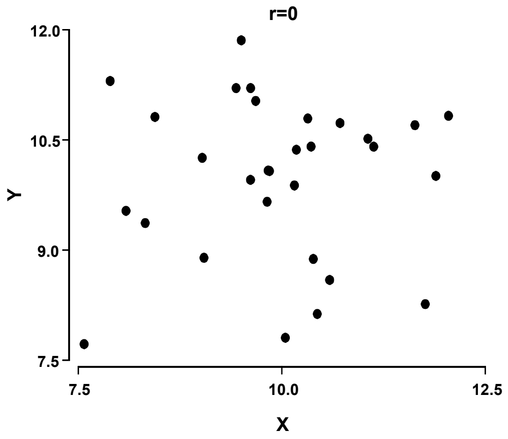
- Value of -1 is as for +1 but line has negative slope

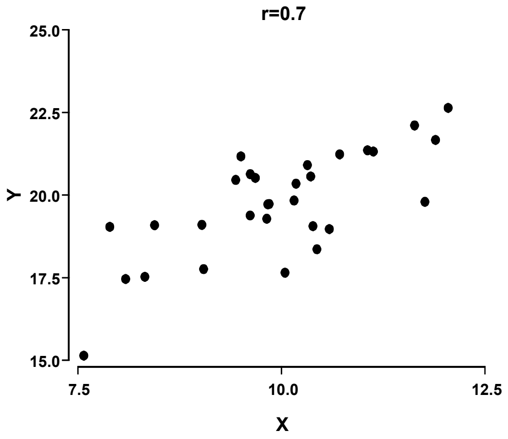
- Value of zero means there is no *linear* relation between the variables

- Other values denote relationships of intermediate strengths

Illustrations



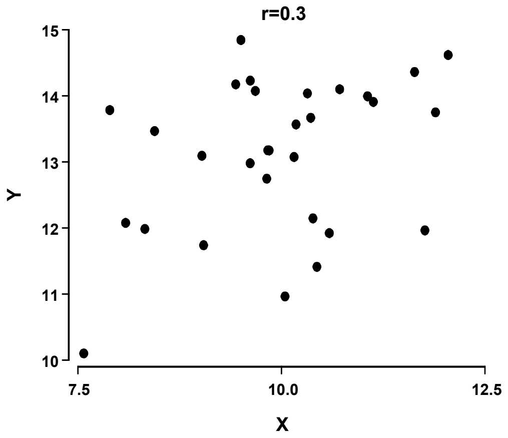




r=0.7 and -0.7 are clearly

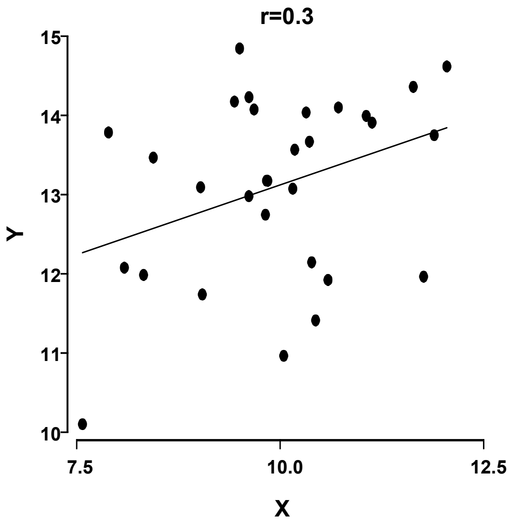
stronger relationships

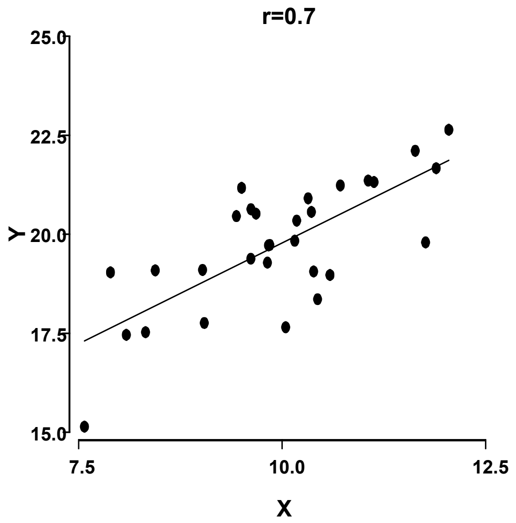
than r=0 or r=0.3



Meaning of “Strength of the Relationship”

Ignore sign of r, this just says whether y tends to increase or decrease with x





Spread about line appears to be greater when r=0.3 than r=0.7.

Approximate Definition of r

Spread of y variable *ignoring* x variable is SD of y, s_y

For FEV1, this is 0.705 l

Spread of y variable after allowing for x is s

For FEV1 and height this is 0.589 l

Proportion of variability of y that remains after *allowing* for x is s/s_y

For FEV1 and height this is 0.836

This is approximately $\sqrt{(1-r^2)}$
(approximate defn. of r)

Giving $r \cong 0.55$ (exact value is $r=0.58$)

$r=\pm1$ means all variation in y is explained by x,
$r=0$ means none of the variation in x is explained by x

Details and Assumptions

Can test hypothesis that true correlation is 0, so you will see statements such as “r=0.35, P=0.07”

(note: P=0.07 doesn’t always go with r=0.35, the P-value depends on r *and* sample size)

Test is exactly the same as test that β=0, so do that instead

This test requires that at least one of x or y is Normal

Can find confidence interval for r, but this requires *both* x and y to be Normal

Drawbacks with Correlation

- Only extreme values easily interpreted
- Attempt to provide single-value summary of relationship between two variables - not really feasible.
- Can be convenient when there are many variables
- Regression usually provides a much more informative analysis

Further reading on regression and correlation: Altman, DG, *Practical Statistics for Medical Research*, Chapman & Hall, 1991, chapter 11; Bland, JM, *An Introduction to Medical Statistics*, OUP, 1995 chapter 11.