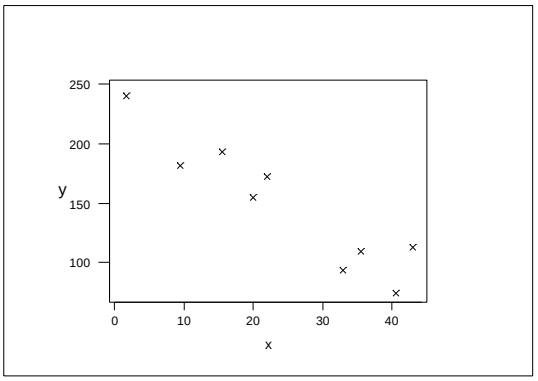
5. Linear regression and correlation

If we measure a response variable y at various values of a controlled variable x, linear regression is the process of fitting a straight line to the mean value of y at each x. For example you might measure fuel efficiency y at various values of an experimentally controlled external temperature x, and then try to fit a straight line to the results (assuming you think there is an underlying linear relationship).



e.g. the plot of *y* against *x* suggests that it is reasonable to fit a straight line.

Model

Say we take *n* measurements of a function $y(x)$, obtaining for each $x_{i\,\mathrm{i}}$ a value y_i. When plotted on a scatter diagram, there is a straight line relationship between y and x, apart from random variation in each y measurement.

Model: $y_i = a + b\,x_i + e_i$

where $a + bx_i$ is the linear relation and e_i is the random error. We assume $e_i \sim N(0, \sigma^2)$ for all *i*, and e_i's are independent, and want to estimate *a* and *b*, using the data.

The likelihood P(D|a,b) can be found since e_i are Normal, i.e. $P(e_i) \propto e^{-\frac{e_i^2}{2\sigma^2}}$, hence since $e_i = y_i - a - b\,x_i$ we have the log likelihood

$$\log P(\{x_i, y_i\}|a, b) = -\sum_i \frac{(y_i - a - bx_i)^2}{2\sigma^2}$$

The maximum likelihood estimator can be found by maximizing this log likelihood. This is equivalent to minimizing

$$E = \sum_i e_i^2 = \sum_i (y_i - a - bx_i)^2$$

since σ^2 is a constant. Minimizing E is minimizing the squared error.

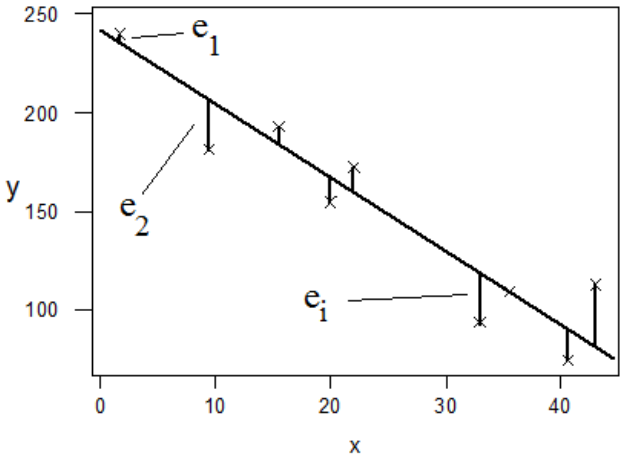
Even when the random error is more complicated than a simple Normal, E can still be defined, and least-squared values can be calculated, though they may not have a very clear interpretation. In fact regression is often used completely blindly, without knowing the model the samples are drawn from, and can still be useful to identify correlations between variables.

Least squares estimates of *a* and *b*

The least-squares estimates $\hat{a}$ and $\hat{b}$ must satisfy $\frac{\partial E}{\partial a} = 0$ and $\frac{\partial E}{\partial b} = 0$, i.e.

(1) $\frac{\partial E}{\partial a} = -2\sum_i (y_i - a - bx_i) = 0$

(2) $\frac{\partial E}{\partial b} = -2\sum_i x_i (y_i - a - bx_i) = 0$



Therefore, $\hat{a}$ and $\hat{b}$ satisfy:

$$(1)\ \sum_i \left(y_i - \hat{a} - \hat{b}\, x_i\right) = n(\bar{y} - \hat{a} - \hat{b}\,\bar{x}) = 0$$

$$(2)\ \sum_i x_i \left(y_i - \hat{a} - \hat{b}\, x_i\right) = \sum_i x_i y_i - n\bar{x}\hat{a} - \hat{b}\sum_i x_i^2 = 0$$

Solving (1) gives: $\hat{a} = \bar{y} - \hat{b}\,\bar{x}$. Substituting into (2) then gives:

$$\sum_i x_i y_i - n\bar{x}(\bar{y} - \hat{b}\,\bar{x}) - \hat{b}\sum_i x_i^2 = \sum_i x_i y_i - n\bar{x}\bar{y} - \hat{b}\left(\sum_i x_i^2 - n\,\bar{x}^2\right) = 0$$

Solving for $\hat{b}$ gives the final answer

$$\hat{b} = \frac{S_{xy}}{S_{xx}} \text{ and } \hat{a} = \bar{y} - \hat{b}\,\bar{x}$$

Here

$$S_{xy} = \sum_i x_i y_i - \frac{\sum_i x_i \sum_i y_i}{n} = \sum_i (x_i - \bar{x})(y_i - \bar{y})$$

$$S_{xx} = \sum_i x_i^2 - \frac{(\sum_i x_i)^2}{n} = \sum_i (x_i - \bar{x})^2$$

It can be shown that $\hat{a}$ and $\hat{b}$ are unbiased. Note S_{xx} is just proportional to the sample variance.

The fitted regression line is:

$$\hat{y} = \hat{a} + \hat{b}x$$
$$= \bar{y} - \hat{b}\bar{x} + \hat{b}x$$

$$\Rightarrow \hat{y} - \bar{y} = \hat{b}(x - \bar{x})$$

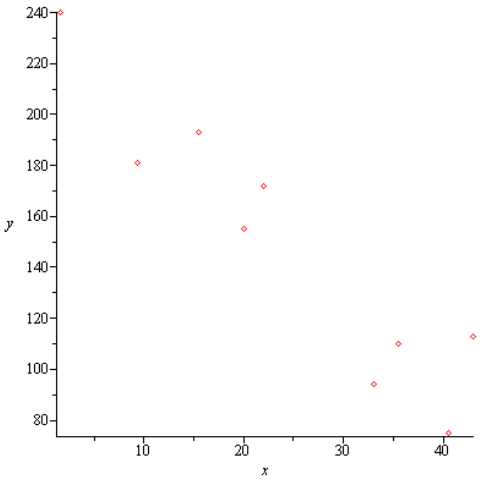
i.e. the regression line passes through $(\bar{x}, \bar{y})$

Example:

The data *y* has been observed for various values of *x*, as follows:

y	240	181	193	155	172	110	113	75	94
x	1.6	9.4	15.5	20.0	22.0	35.5	43.0	40.5	33.0

Fit the simple linear regression model using least squares.



Answer :

$n = 9$
$\sum_i x_i = 220.5\ ,\ \sum_i y_i = 1333.0,$
$\sum_i x_i^2 = 7053.7,\ \sum_i x_i y_i = 26864, \sum_i y_i^2 = 220549$

$$S_{xy} = 26864 - \frac{220.50 \times 1333.0}{9} = -5794.1$$

$$S_{xx} = 7053.7 - \frac{220.5^2}{9} = 1651.42$$

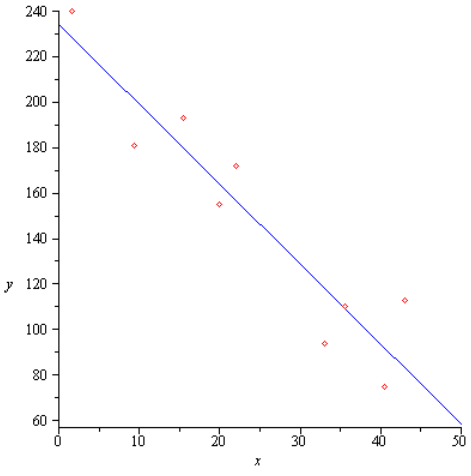
Hence

$$\hat{b} = \frac{S_{xy}}{S_{xx}} = -\frac{5794.5}{1651.45} = -3.5086$$

$$\hat{a} = \bar{y} - \hat{b}\bar{x} = \frac{1333.0}{9} - (-3.5086) \times \frac{(220.50)}{9} = 234.1$$

So the fit is approximately

$$y = 234.1 - 3.509x$$



Estimating σ^2: variance of *y* about the fitted line

Estimated error is: $\hat{e}_i = y_i - \hat{y}_i$

The mean error is zero, so the ordinary sample variance of the e_i's is $\sim \frac{1}{n-1}\sum_i \hat{e}_i^2$

In fact, this is biased since two parameters, *a* and *b* have been estimated. The unbiased estimate is:

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum \hat{e}_i^2 = \frac{1}{n-2}\sum (y_i - \hat{y}_i)^2$$

Using $\hat{y}_i = \bar{y} + \hat{b}(x_i - \bar{x})$ then

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum \left((y_i - \bar{y}) - \hat{b}(x_i - \bar{x})\right)^2 = \frac{1}{n-2}\sum \left((y_i - \bar{y})^2 - 2\,(y_i - \bar{y})(x_i - \bar{x}) + \hat{b}^2(x_i - \bar{x})^2\right)$$
$$= \frac{1}{n-2}\left(S_{yy} - 2\,\hat{b}S_{xy} + \hat{b}^2 S_{xx}\right) = \frac{1}{n-2}\left(S_{yy} - 2\,\hat{b}S_{xy} + \hat{b}\frac{S_{xy}}{S_{xx}}S_{xx}\right) = \frac{S_{yy} - \hat{b}S_{xy}}{n-2}$$

Confidence interval for the slope, b

Recall that for Normal data with unknown variance, confidence interval for μ is:

$$\bar{X} - t_{n-1}\sqrt{\frac{s^2}{n}} \quad \text{to} \quad \bar{X} + t_{n-1}\sqrt{\frac{s^2}{n}}$$

The quantity s^2/n is the estimate of σ^2/n, the variance of $\bar{X}$

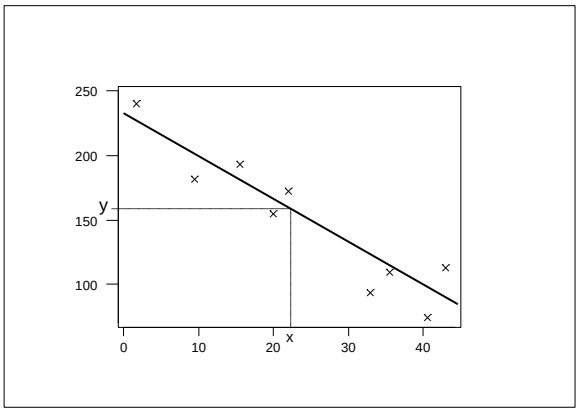
It can be shown that var($\hat{b}$) = σ^2/S_{xx}, estimated by $\hat{\sigma}^2/S_{xx}$ (*n*-2 df).

Confidence interval for *b* is $\hat{b} - t_{n-2}\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}$ to $\hat{b} + t_{n-2}\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}$

Predictions

One common reason for fitting a linear regression model is to predict y given x.

Predicted value for the mean y at x is $\hat{y} = \hat{a} + \hat{b}x$.

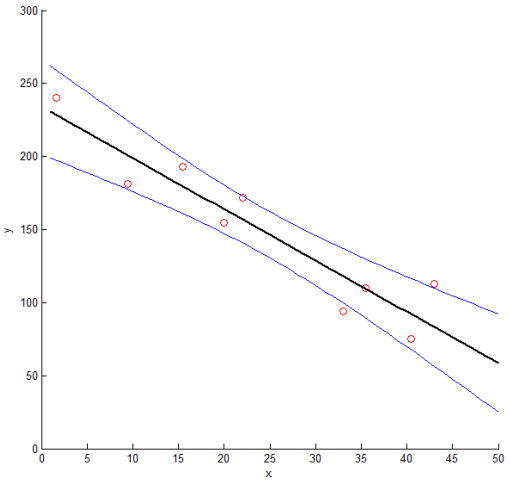


Confidence interval for mean *y* at given *x*

It can be shown that
var($\hat{y}|x$) = var($\hat{a} + \hat{b}x$) =

$$\sigma^2\left(\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}\right)$$

Therefore, confidence interval for mean y is: $\hat{y}_i \pm t_{n-2}\sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}\right)}$



Extrapolation: using the fitted model to predict y outside the range of *x*'s used estimating *a* and *b*. This may be misleading, as the approximate linear relation may not continue to hold beyond the range for which you have observations.

Example: Using the previous data, what is the mean value of y at $x = 30$ and the 95% confidence interval?

Answer

The expected value is $\hat{y} = \hat{a} + \hat{b}x = 234 - 3.51 \times 30 = 128.8$

For the confidence interval need (with $n = 9$)

$$\hat{\sigma}^2 = \frac{S_{yy} - \hat{b}S_{xy}}{n-2} = 398.28$$

For 95% confidence need $t_{n-2} = t_7$ for Q=0.975, i.e. $t_7 = 2.3646$.

Hence confidence interval is

$$\hat{y} \pm t_7 \sqrt{\hat{\sigma}^2 \left(\frac{1}{n} + \frac{(30 - \bar{x})^2}{S_{xx}} \right)} = 128.8 \pm 2.3646 \sqrt{398.28 \left(\frac{1}{9} + \frac{\left(30 - \frac{220.5}{9}\right)^2}{1651.42} \right)}$$

$$\approx 129 \pm 17$$

Confidence interval for a prediction

Often we want to predict the range a future data point might lie, rather than just calculate the mean. This confidence interval for a single response (measurement of y at x_0) is given by

$$\hat{a} + \hat{b}x_0 \pm t_{n-2} \sqrt{\hat{\sigma}^2 \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right)}$$

This is larger because it is a combination of the uncertainty in the mean, and the expected scatter of a given point about the mean.

Example: Using the previous data, what is the 95% confidence interval for a measurement of y at $x = 30$?

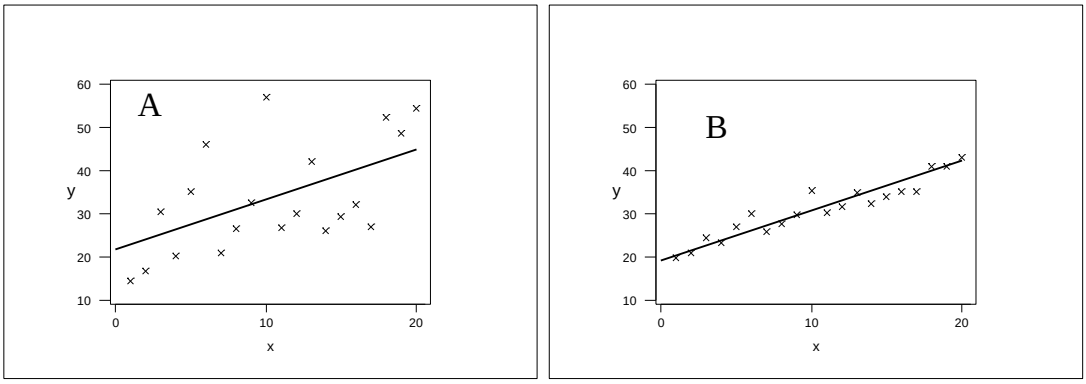
Answer

$$\hat{y} \pm t_7 \sqrt{\hat{\sigma}^2 \left(1 + \frac{1}{n} + \frac{(30 - \bar{x})^2}{S_{xx}} \right)} = 129 \pm 2.34 \sqrt{398.28 \left(1 + \frac{1}{9} + \frac{\left(30 - \frac{220.5}{9}\right)^2}{1651.42} \right)}$$

$$\approx 129 \pm 50$$

Correlation

Regression tries to model the relation between *y* and *x*. Correlation tries to measure the strength of the linear association between *y* and *x*.



Same fitted line in both cases, but stronger linear association in case B.

What does correlation mean? If *x* and *y* are positively correlated, then if *x* is high *y* is expected to be high, if *x* is low then *y* is expected to be low. In other words, on average $(x - \bar{x})(y - \bar{y})$ is expected to be positive: both if *x* and *y* are below the mean, or if *x* and *y* are above the mean. Similarly for a negative correlation $(x - \bar{x})(y - \bar{y})$ is expected to be negative.

We can therefore use $S_{xy} = \sum_i (x_i - \bar{x})(y_i - \bar{y})$ to quantify the correlation. It is often convenient to normalize by the variance of the *x* and *y*, giving the definition of the correlation coefficient:

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$$

This is sometimes called the Pearson product-moment. The range is: $-1 \leq r \leq 1$:

r = 1: there is a line with positive slope going through all the points;

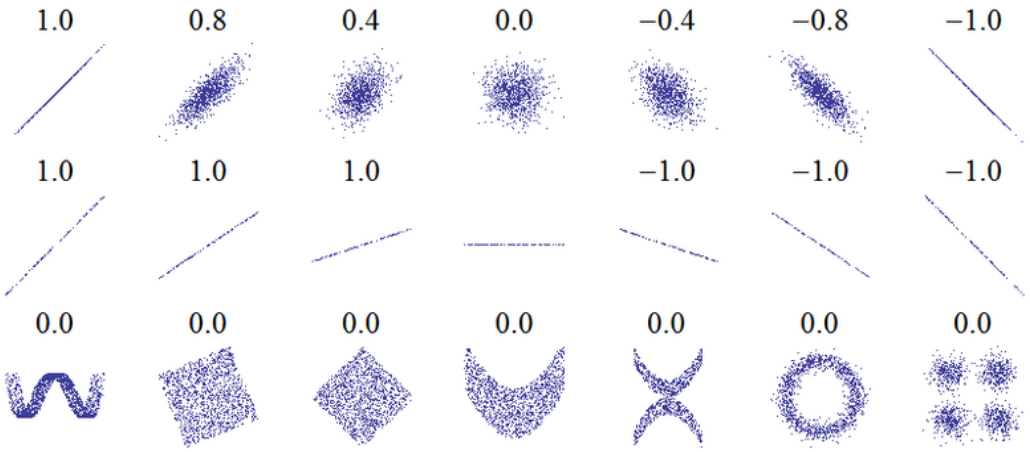
r = -1: there is a line with negative slope going through all the points;

r = 0: there is no linear association between *y* and *x*.

Example: from the previous data, $S_{xy} = -5794, S_{xx} = 1651, S_{yy} = 23117$ hence

$$r = \frac{-5794}{\sqrt{1651 \times 23117}} \approx -0.94$$

The magnitude of *r* measures how noisy the data is, but not the slope. Also $r = 0$ only means that there is no linear relationship, and does not imply the variables are independent – there could be many more complicated relationships that do not fit a straight line:



In general it is not easy to quantify the error on the estimated correlation coefficient. Possibilities include subdividing the points and assessing the spread in r values.

Also, of course $r \neq 0$ does not imply that changes in x *cause* changes in y - additional types of evidence are needed.

Example: Earnings and height:

Log earnings and height

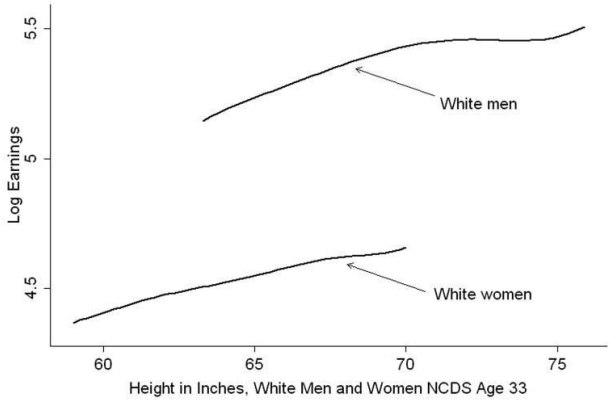
error

Correlation r

	Men		Women	
Dependent variable:	**Height coefficient**	**Number of observations**	**Height coefficient**	**Number of observations**
		NCDS		
Log weekly gross earnings	0.026	4,927	0.024	5,033
	(0.004)		(0.007)	
Log average hourly gross earnings	0.023	4,860	0.019	4,995
	(0.004)		(0.005)	
		BCS		
Log weekly gross earnings	0.014	2,265	0.029	2,136
	(0.003)		(0.006)	
Log average hourly gross earnings	0.010	2,253	0.015	2,127
	(0.003)		(0.004)	
		PSID		
Log weekly earnings	0.023	23,465	0.014	21,271
	(0.004)		(0.006)	
Log average hourly earnings	0.019	23,465	0.012	21,271
	(0.004)		(0.003)	

Notes. OLS regression coefficients reported for height in inches, with standard errors in parentheses. The NCDS and PSID regressions use multiple observations per person, and unobservables are clustered at the individual level. The NCDS and BCS samples are restricted to those for whom we have test scores at ages 7 and 11 (NCDS), or 5 and 10 (BCS). The PSID sample consists of white household heads or wives between the ages of 25 and 60, inclusive, between 1988 and 1997. NCDS and BCS regressions include indicators for ethnicity, and the NCDS regressions also include an age indicator. The PSID regressions include a set of age and year indicators.

J Polit Econ. 2008; 116(3): 499–532.
http://www.journals.uchicago.edu/doi/abs/10.1086/589524



So there is strong evidence for a 2-3% correlation. This doesn't mean being tall *causes* you earn more (though it could). For example height could be correlated with cognitive ability, and cognitive ability causes you to earn more. In fact this appears to be the case: height is correlated with intelligence, both higher height and higher intelligence being caused by better health and nutrition during development. There could also be a genetic component (maybe smart women slightly prefer tall men – perhaps because it is an indicator of health and nutrion – and they then have tall smart children). Determining the the reason for an empirical correlation is usually extremely difficult.