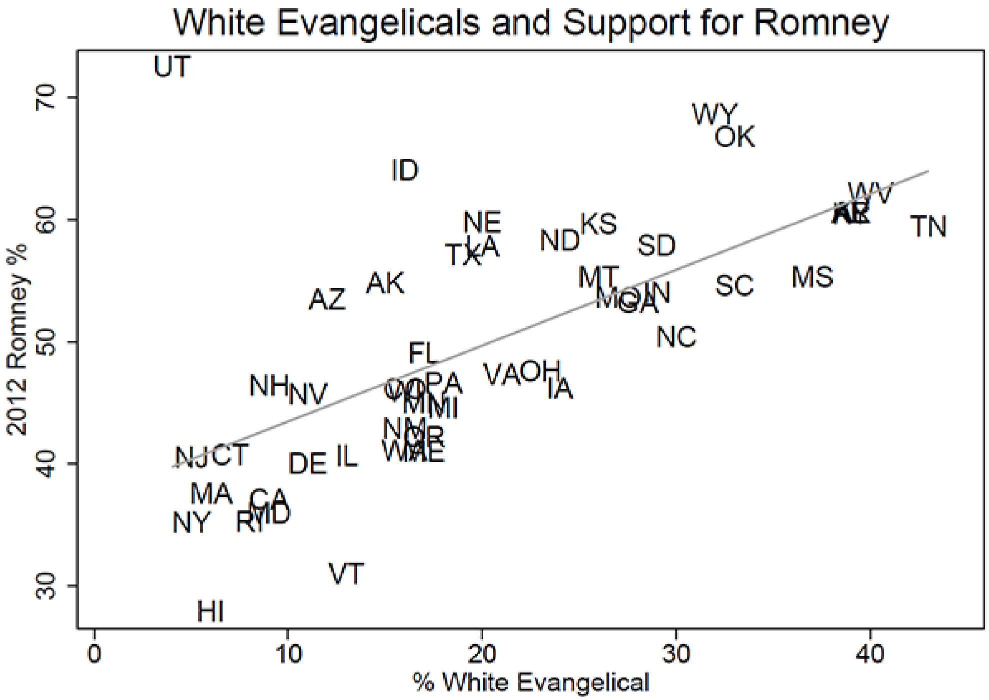
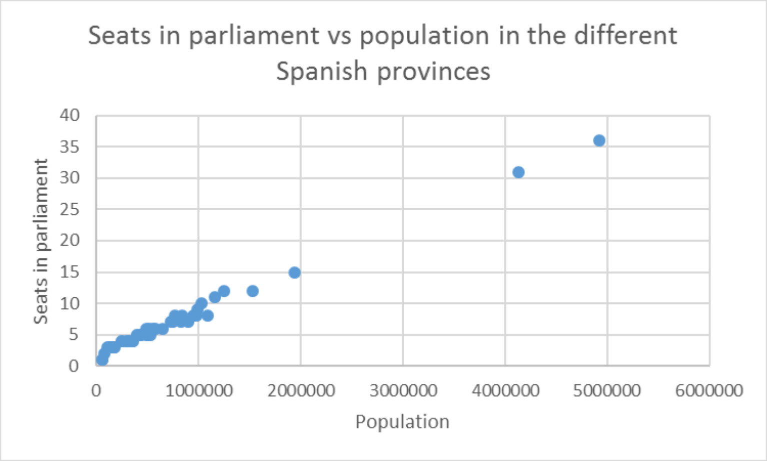
Class 7: Covariance, correlation and regression



Objective



We saw in the last class how seats in parliament is approximately linearly related to population. Now we want a measure of how close a relationship is to linear.

Then, we want to fit the “best” line possible to the data.

Finally, we want to check whether the fit is reasonable.

Covariance

The covariance is designed as a measure of whether the relationship between two variables is positive (generally increasing) or negative (generally decreasing).

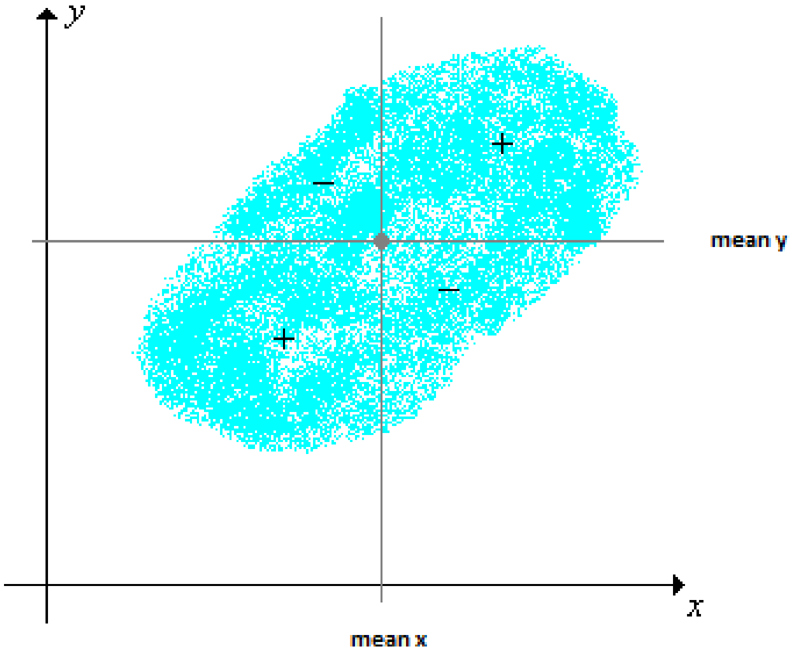
For data (x_1,y_1), …, (x_N, y_N), the covariance is:

$$\hat{\sigma}_{xy} = \frac{1}{N}\{(x_1 - \bar{x})(y_1 - \bar{y}) + \cdots + (x_N - \bar{x})(y_N - \bar{y})\}$$

The quasi covariance is:

$$s_{xy} = \frac{1}{N-1}\{(x_1 - \bar{x})(y_1 - \bar{y}) + \cdots + (x_N - \bar{x})(y_N - \bar{y})\}$$

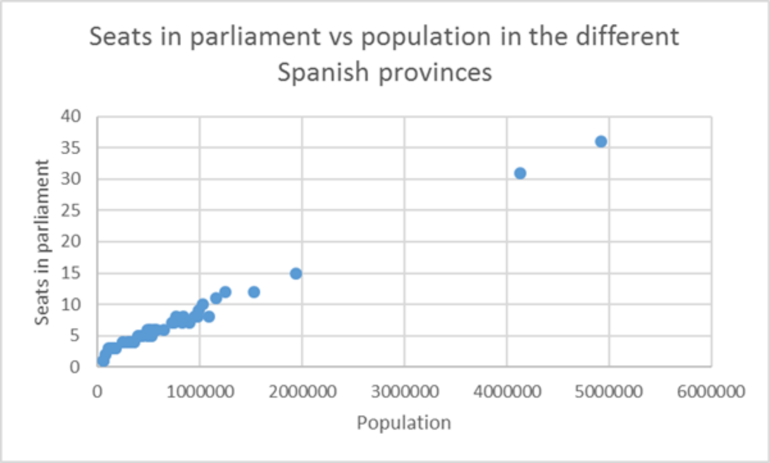
Covariance



For an increasing relationship, most data are in the top right and bottom left regions so the covariance is positive.

For a decreasing relationship, the covariance will be negative.

Covariance

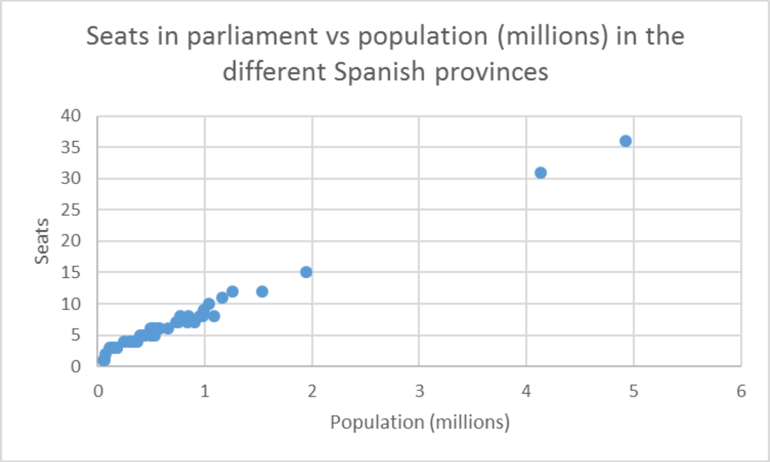


The covariance is $\hat{\sigma}_{xy} \approx$ 5.2 millions.

Does this mean we have a very strong relationship?

What would happen if we measured population in millions?

Covariance



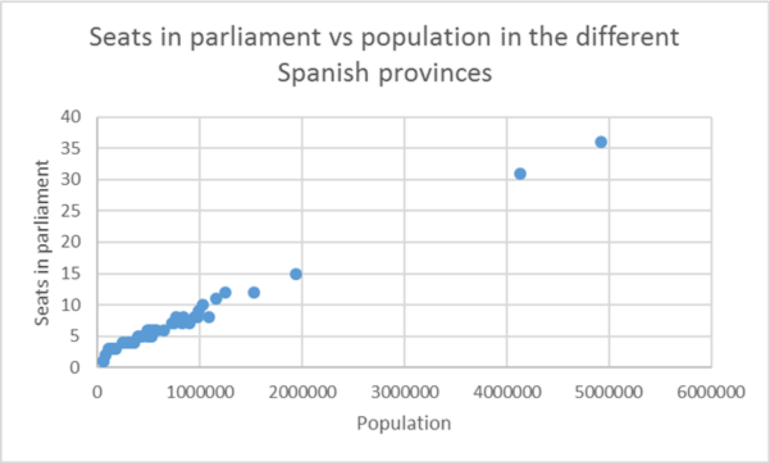
The relationship between population and seats in parliament looks the same but the covariance is now $\hat{\sigma}_{xy} \approx 5.2$.

We need an alternative measure which doesn’t depend on the units of the data.

Correlation

The correlation is $r_{xy} = \frac{\hat{\sigma}_{xy}}{\hat{\sigma}_x \hat{\sigma}_x} = \frac{s_{xy}}{s_x s_y}$.

This is the covariance divided by the product of the standard deviations.



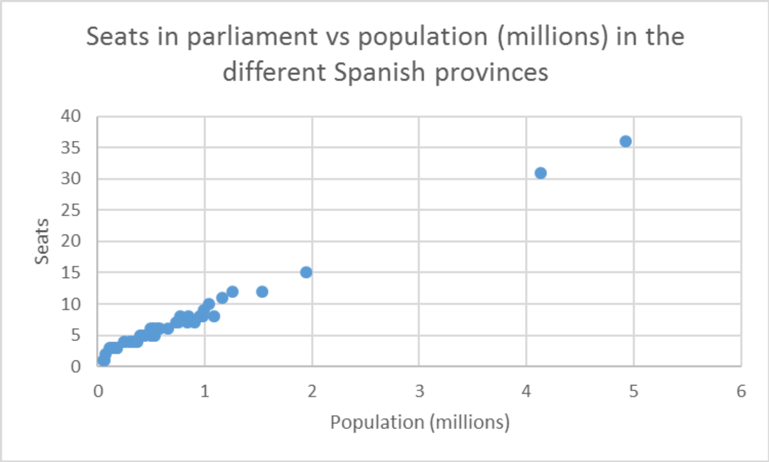
Does this depend on the units?

r_{xy} = 0.995.

Correlation

The correlation is $r_{xy} = \dfrac{\hat{\sigma}_{xy}}{\hat{\sigma}_x \hat{\sigma}_x} = \dfrac{s_{xy}}{s_x s_y}$.

This is the covariance divided by the product of the standard deviations.



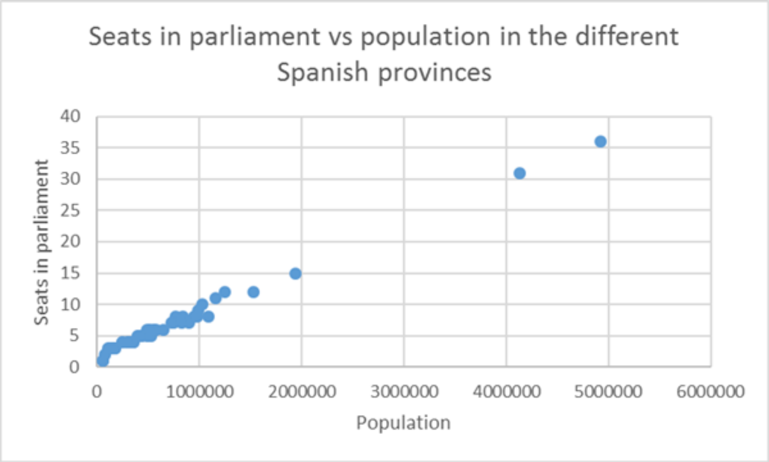
The correlation is still the same.

r_{xy} = 0.995.

How do we interpret this number?

Properties of the correlation coefficient

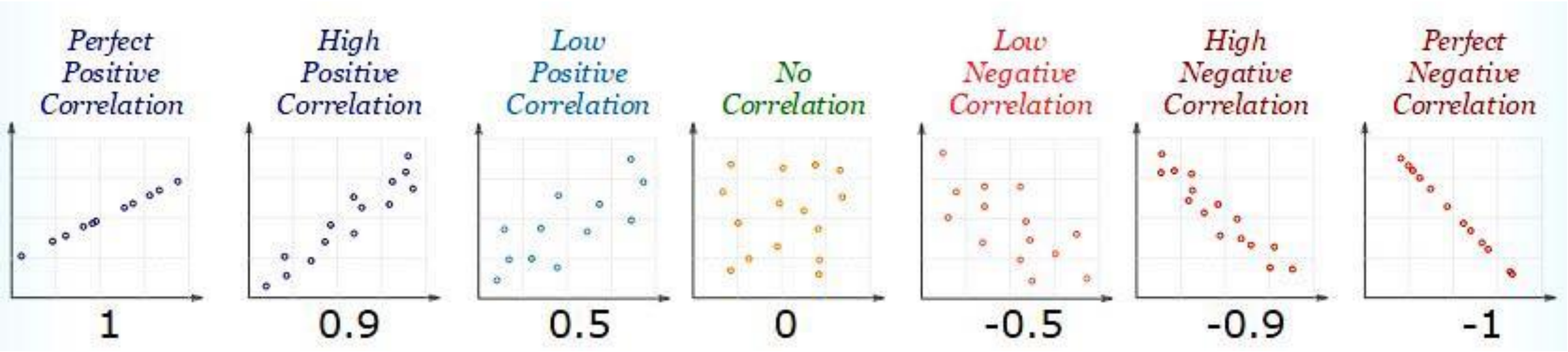
- $-1 \leq r_{xy} \leq 1.$
- If $r_{xy} = 1$, then x and y follow an exact, increasing or positive, linear relationship.
- If $r_{xy} = -1$, then x and y follow an exact, decreasing, or negative linear relationship.
- If there is no relation between the two variables, then $r_{xy} = 0$.
- The closer the correlation is to 1 (or -1), the closer the data are to a line.



r_{xy} = 0.995.

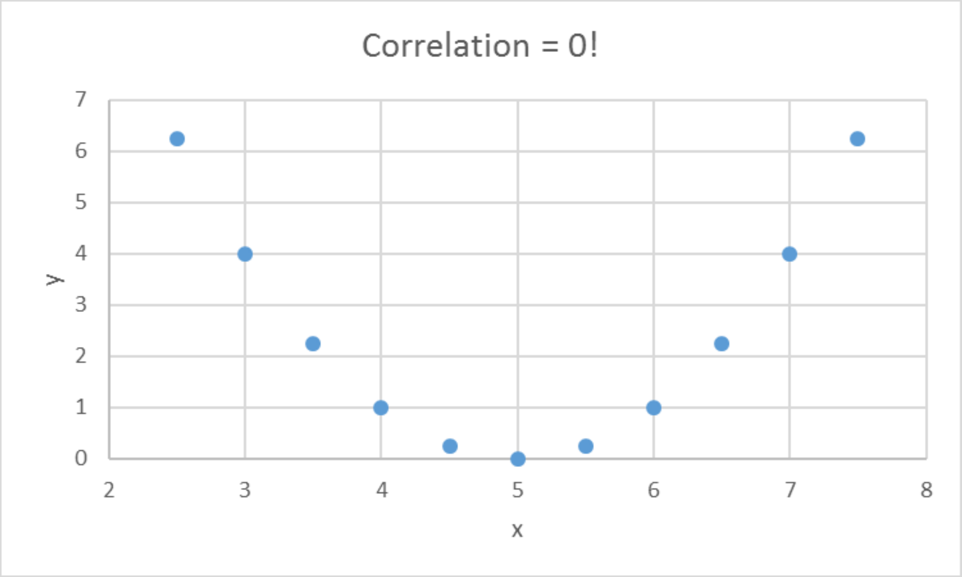
The data are very close to a line.

Examples: different levels of correlation



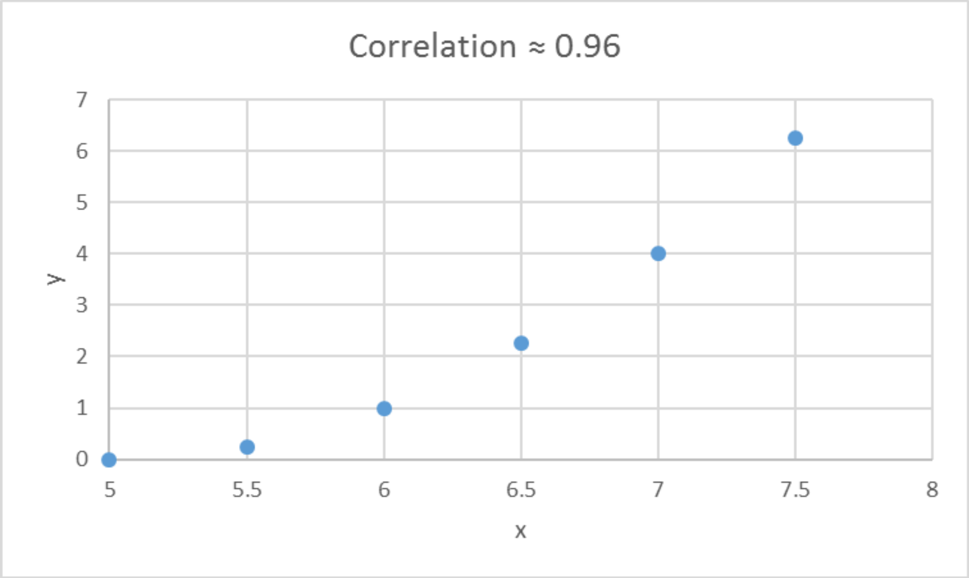
As the correlation gets closer to zero, the data are more spread out from a line.

Example: zero correlation doesn’t mean no relationship



Overall, the relationship is neither increasing nor decreasing.

Example: high correlation doesn't mean a line fit always looks good



Overall, the points are close to a line, but these data are just the right hand side of the previous slide.

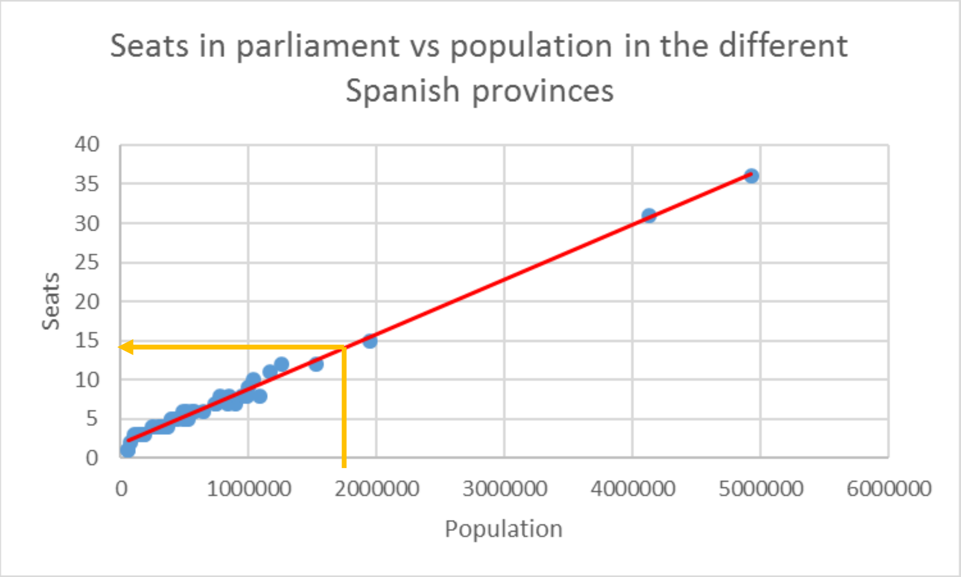
Before trying to fit a regression line, we should always look at the data to see if this is appropriate.

Regression

Often we want to predict the values of y in terms of x.

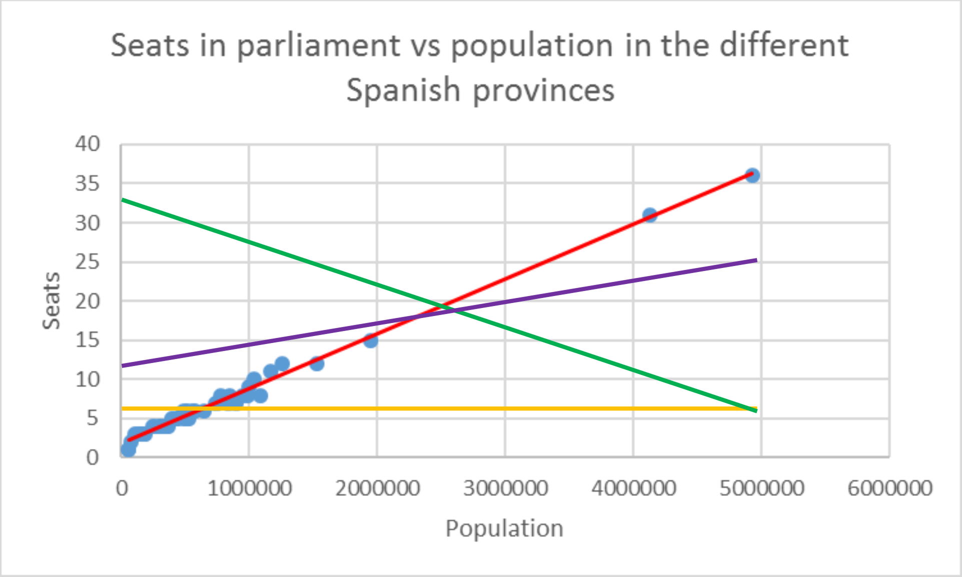
If Spain decides to annex the district of Porto and turn this into a Spanish province, how many seats in parliament should it be assigned?

If we fit a line $y = \alpha + \beta\, x$ to the data we have, then we can make the prediction by letting x be the population of Porto and calculating y.



Which regression line should we choose?

We could fit many different lines to the data.

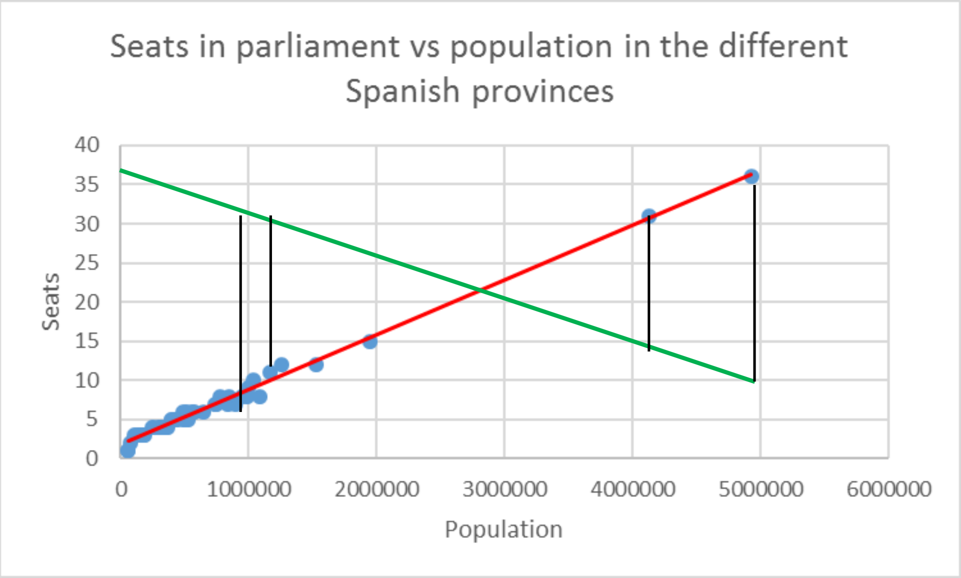


We need to compare how well each line fits and select the best one.

Choosing a regression line

For the data we have: $(x_1,y_1), \dots, (x_N, y_N)$, if we fit a line $y = a + b x$, then the errors or residuals are

$r_1 = y_1 - (a+bx_1), \dots, r_N = y_N - (a+bx_N)$,



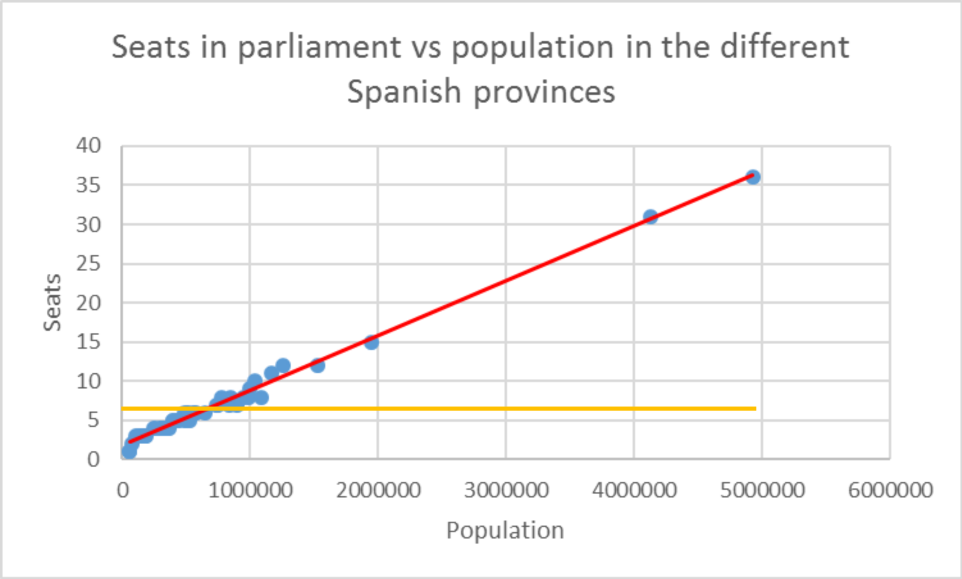
Here we have some of the residuals for the Green line.

The residuals for the red line are much smaller.

Choosing a regression line

We want to make the overall error as small as posible.

We could try to choose a line so that the sum of the residuals is equal to zero.



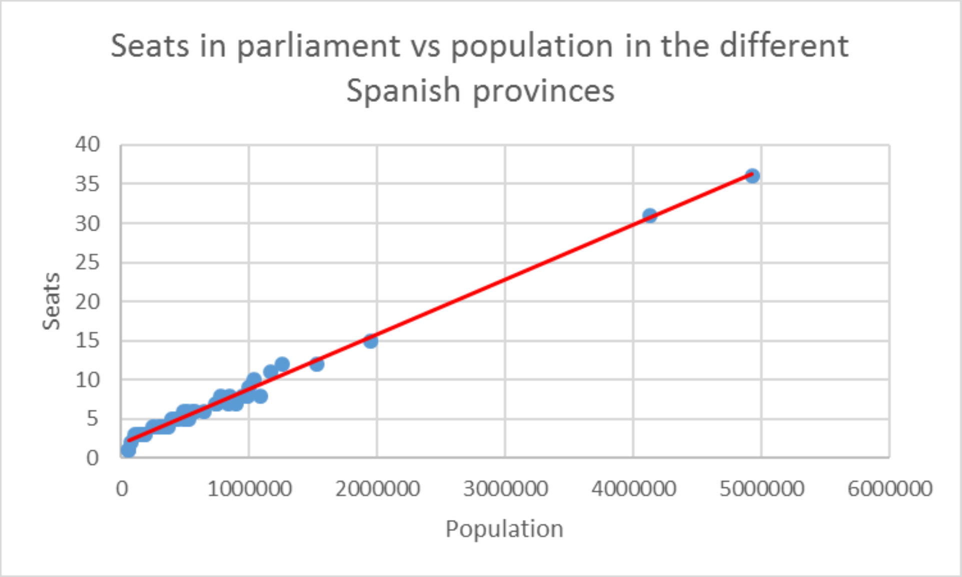
The red line and the yellow line at the mean value of Seats satisfy this.

The problem is some residuals are negative and some positive.

Least squares

We could also choose to minimize the sum of the absolute values of the errors. (Bad statistical properties)

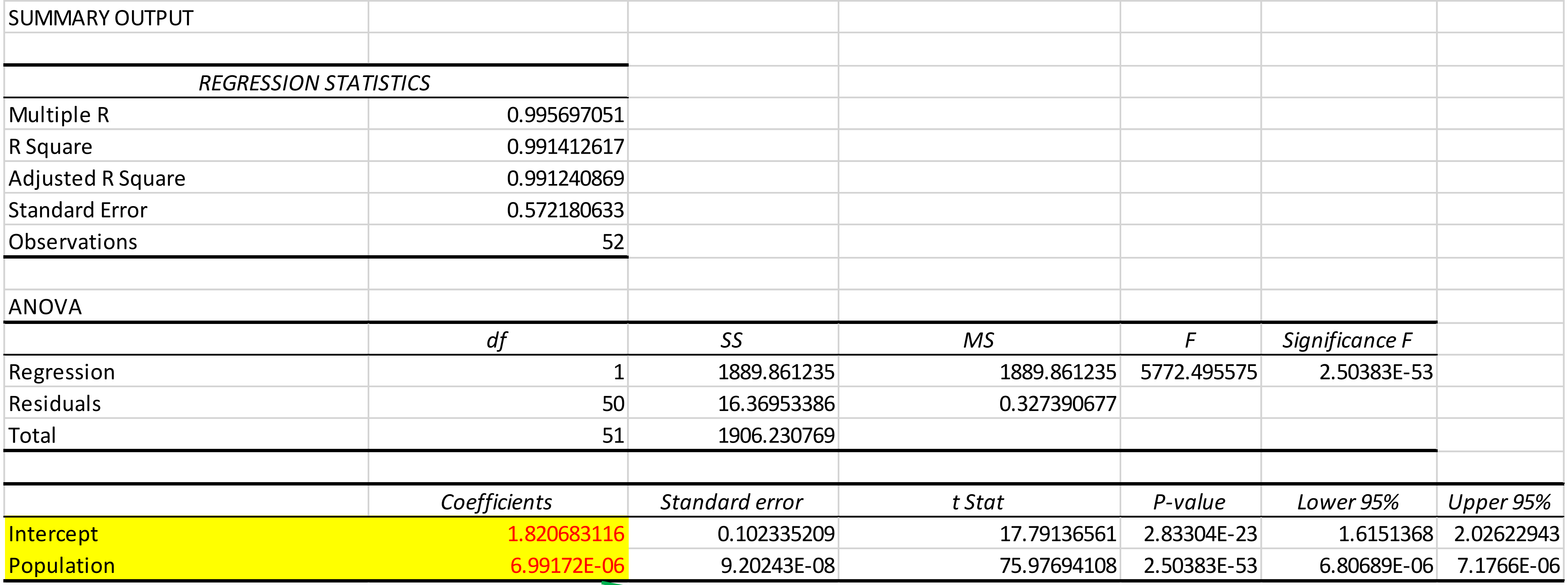
Instead we choose the line to minimize the sum of squared errors: $r_1^2 + \ldots + r_N^2$.



Does this idea remind you of a measure of error we have seen before?

The red line is the least squares fit..

Excel output



SUMMARY OUTPUT	
REGRESSION STATISTICS	
Multiple R	0.995697051
R Square	0.991412617
Adjusted R Square	0.991240869
Standard Error	0.572180633
Observations	52

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	1889.861235	1889.861235	5772.495575	2.50383E-53
Residuals	50	16.36953386	0.327390677		
Total	51	1906.230769			

	Coefficients	*Standard error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	1.820683116	0.102335209	17.79136561	2.83304E-23	1.6151368	2.02622943
Population	6.99172E-06	9.20243E-08	75.97694108	2.50383E-53	6.80689E-06	7.1766E-06

This is scientific notation. E-06 means move the decimal point 6 places to the left.

The line is Seats = 1.821 + 0.00000699 x Population

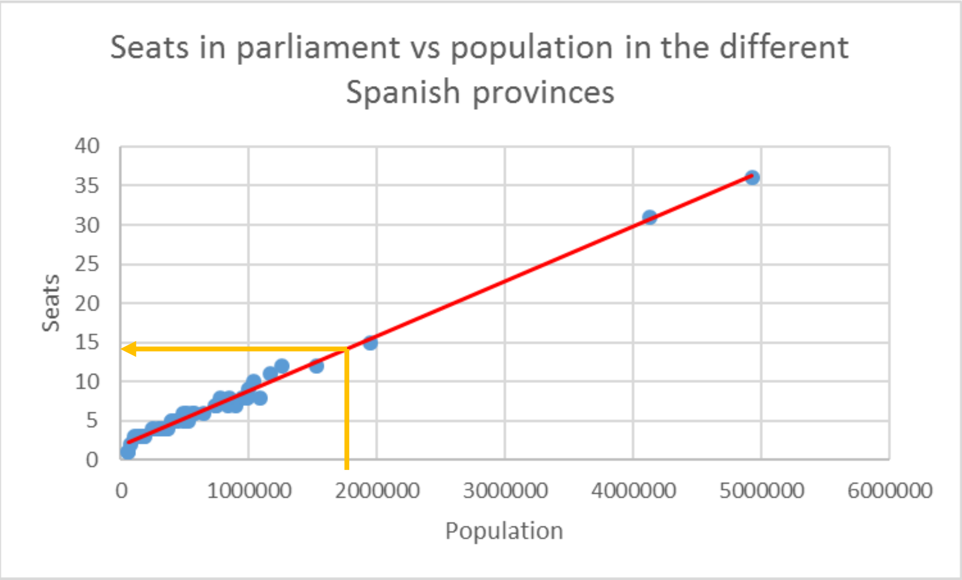
Prediction

The line is: Seats = 1.821 + 0.00000699 x Population.

The district of Porto has a population of 1,781,826.

Therefore we would estimate that Porto should have approximately

1.821 + 0.00000699 x 1,781,826 = 14.28 seats in parliament



Prediction

What if we wanted to make Perejil into a province?

No-one lives there, except a few goats!

Using the line: 1.821 + 0.00000699 x 0 = 1.821 seats in parliament.

This doesn’t seem reasonable!

Predictions tend to be good within the data range and get worse the further away we are.

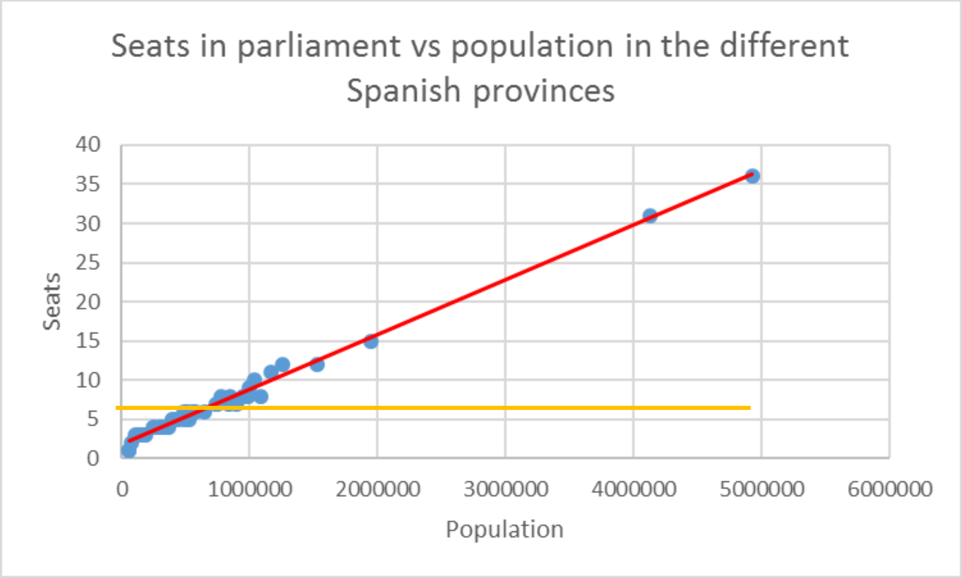
How well is our regression doing: the coefficient of determination

REGRESSION STATISTICS	
Multiple R	0.995697051
R Square	0.991412617
Adjusted R Square	0.991240869
Standard Error	0.572180633
Observations	52

The coefficient of determination is 99.1%:

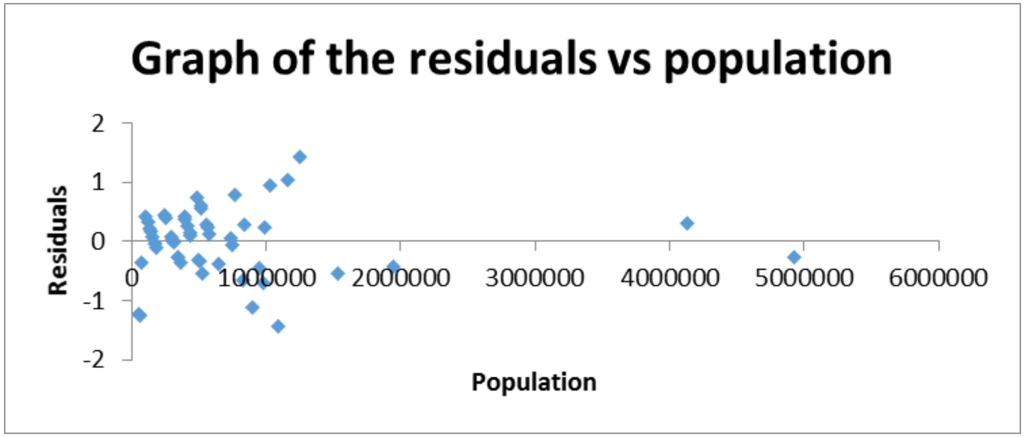
the correlation squared x 100%

This measures how much better our predictions are when we use regression line than when we use the global mean.



How well is our regression doing: residuals

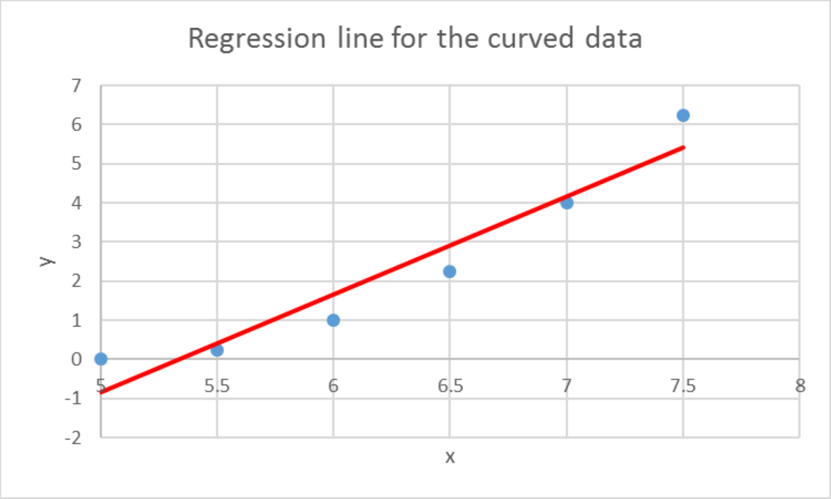
It is sensible to graph the residuals for our fitted regression.



If everything is ok, the residuals should look fairly "random".

What do you think?

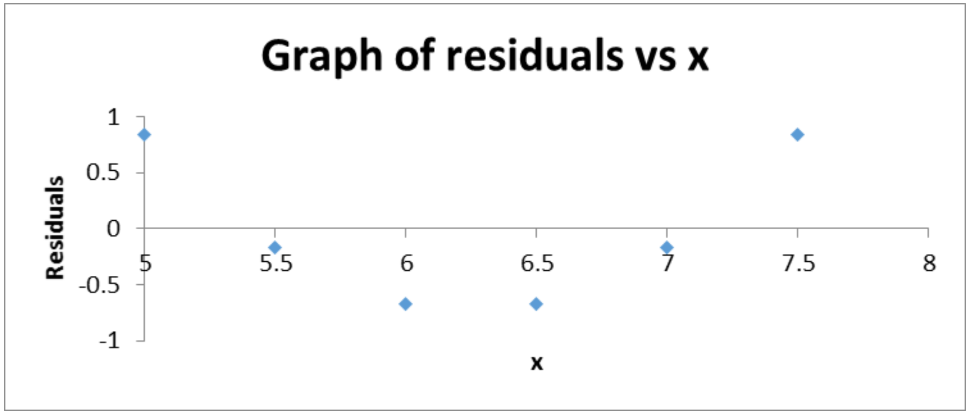
How well is our regression doing: residuals



Remember the curve data. The line is quite close to the points.

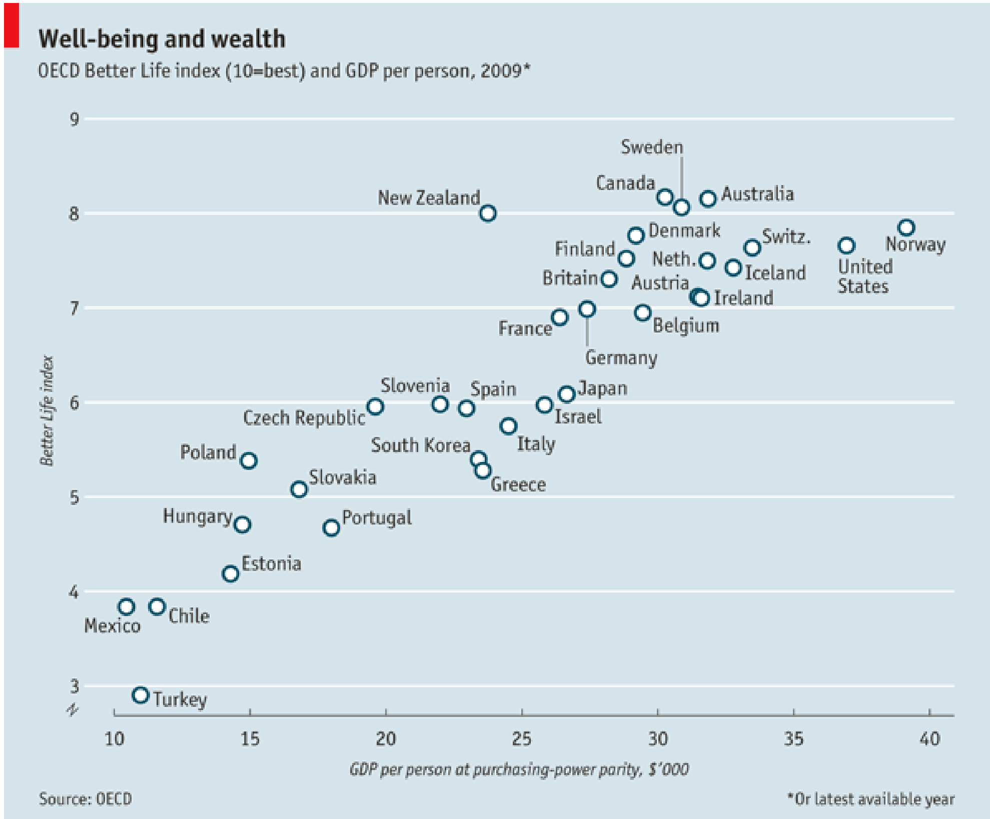
You can see a clear pattern in the residuals.

The simple regression is not a good fit.



Exercise

The following graphic compares levels of well-being (according to the Better Life Index) and wealth (according to GDP per person) in a number of different countries.



Well-being and wealth

OECD Better Life index (10=best) and GDP per person, 2009*

Source: OECD

*Or latest available year

In this case:

a) The correlation between well-being and wealth is positive and the interquartile range of the wealth data is approximately 30.
b) The correlation between well-being and wealth is zero and the range of the well-being data is approximately 3.
c) The correlation between well-being and wealth is negative and the range of wealth data is approximately 30.
d) None of the above.

Exercise

The following graphic compares levels of well-being (according to the 2017 Sustainable Economic Development Assessment or SEDA score) and happiness level (according to the 2017 World Happiness Report) in a number of different countries.

In this case:

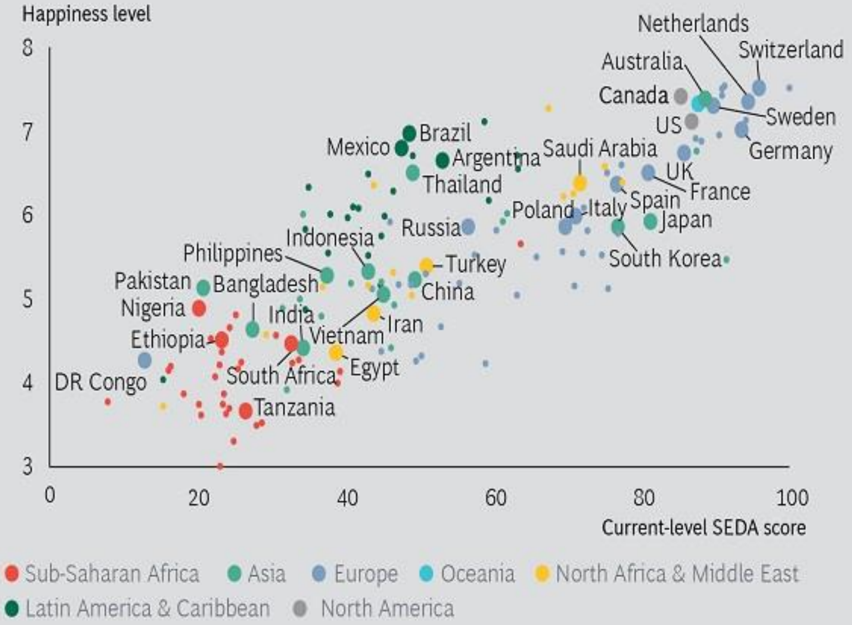
a) The correlation between the SEDA score and the happiness level is positive and close to 1 and the intercept of the regression line relating happiness to SEDA score is negative.

b) The covariance between the SEDA score and the happiness level is positive and the standard deviation of the happiness level is bigger than 4.

c) The interquartile range of the SEDA scores is bigger than 90 and the slope of the regression line relating happiness to the SEDA score is positive.

d) None of the above.

Happiness and Well-Being Are Strongly Linked, with Marked Regional Differences

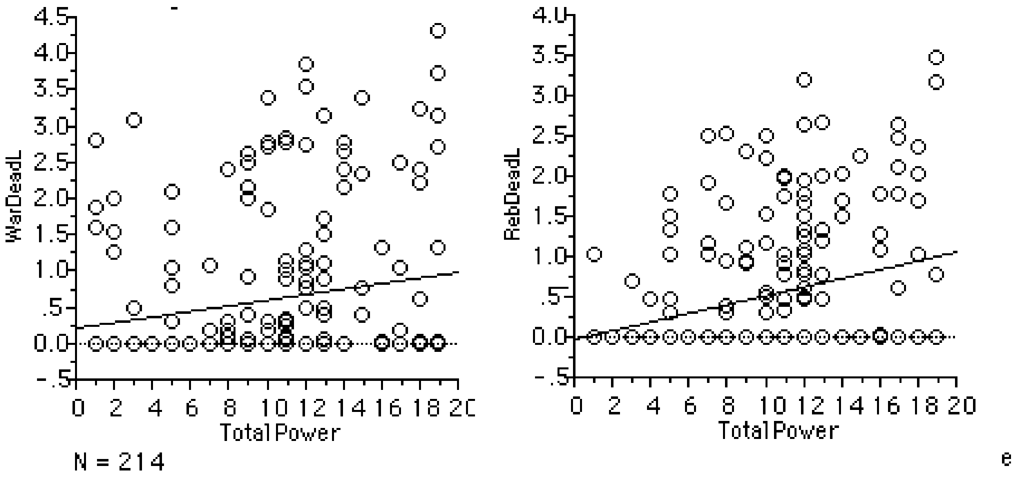


Sources: SEDA 2017; UN 2017 World Happiness Report.
Note: The named countries constitute the 35 countries in our subset with the largest populations and/or the largest economies.

Exercise

The graph below is taken from *The Statistics of Democide*, by R.J. Rummell.

FIGURE 21.2
Scatter Plots of War-Dead and
Rebellion-Dead Against Power



Mark the correct response:

a) The correlation between Total Power and War Dead is positive and close to 1.

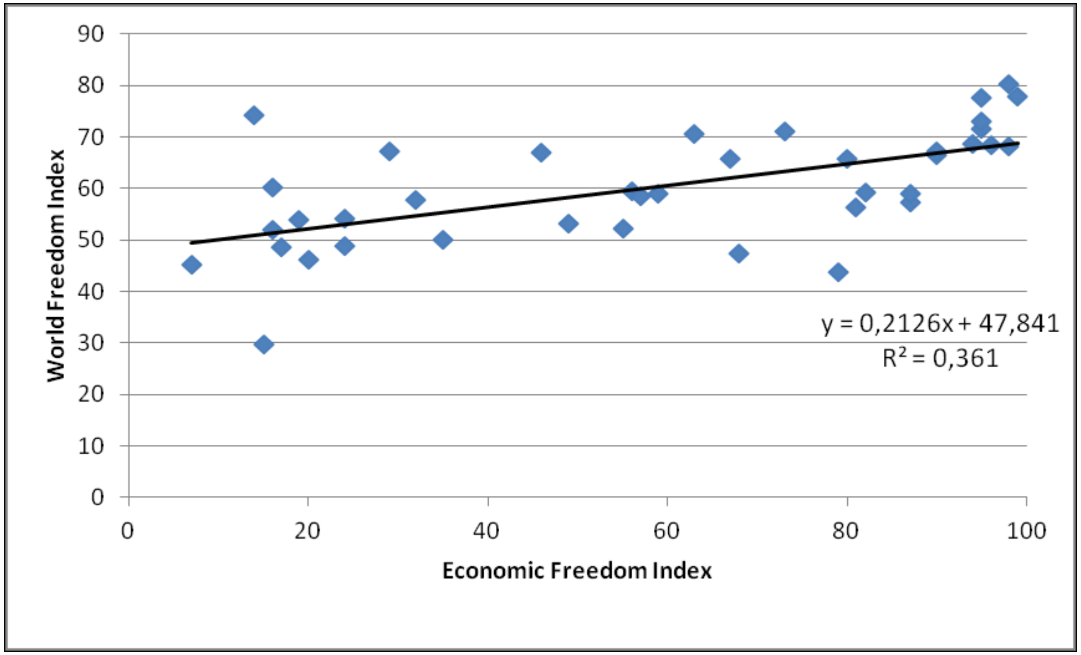
b) The correlation is exactly zero.

c) The correlation is negative.

d) The correlation is positive but close to zero.

Exercise

The following graph relates the 2016 World Freedom Index (WFI) to the 2016 Economic Freedom Index (EFI) for alphabetically chosen countries of the world from A to C.



The Ivory Coast (Côte d'Ivoire) was included in A-C of the EFI with value 60 but not in the WFI (where it was classified as beginning with I). Use the regression line to estimate its WFI value. Do you think the estimate is reasonable? Why?