

Correlation and regression

1. Pearson product moment coefficient of correlation

If we wish to measure the strength of linear relationship between two variables (x and y), and the data is cardinal, we use the Pearson product moment coefficient of correlation (r), which is defined as follows:

$$r = \frac{n \sum xy - \sum x \sum y}{\sqrt{(n \sum x^2 - (\sum x)^2) \times (n \sum y^2 - (\sum y)^2)}}$$

The value of r lies in the range $-1 < r < +1$. A r value of -1 signifies a perfect negative linear correlation, +1 a perfect positive linear correlation, and 0 no linear correlation. These scenarios are most easily demonstrated on a scatter graph. Note that the coefficient only measures the strength of relationship; it is not evidence of cause and effect.

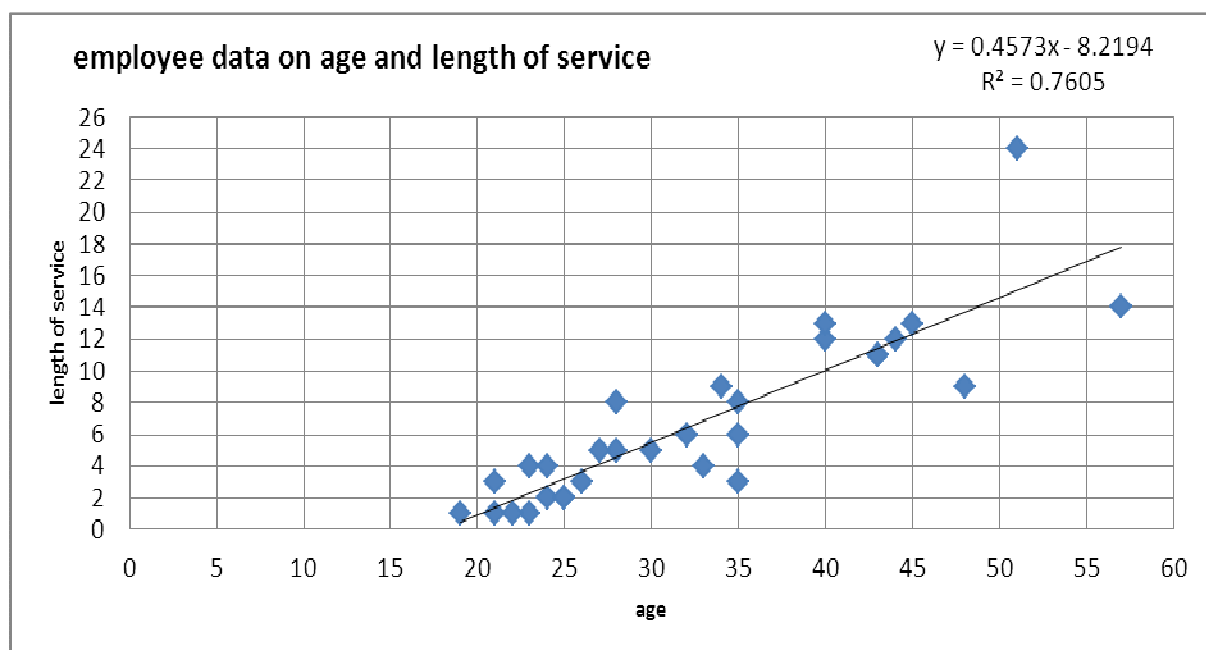
In effect r measures how adequately a scatter of observations can be represented by a straight line. However it is more meaningful to interpret the strength of the relationship by looking at r^2 (known as the coefficient of determination, where $0 < r^2 < +1$) which measures the proportion (or percentage) of variation in y which is explained by the variation in x. To focus on r tends to overstate the strength of the relationship (e.g. $r = 0.7$ seems to suggest a fairly strong relationship, but it explains less than 50% of the variation). We can test the significance of the coefficient using the test statistic:

$$t = \frac{r\sqrt{(N-2)}}{\sqrt{(1-r^2)}}$$

which has a t-distribution with N-2 degrees of freedom

Example

In a survey of 30 company employees, the correlation between length of service and age is 0.87207. A scatter diagram of the data is shown below.



The details of the calculation for the value of the correlation coefficient are shown below:

totals	service	age			
	y	x	y ²	x ²	xy
	2	24	4	576	48
	24	51	576	2601	1224
	2	25	4	625	50
	9	34	81	1156	306
	12	40	144	1600	480
	6	32	36	1024	192
	1	23	1	529	23
	5	28	25	784	140
	3	21	9	441	63
	4	33	16	1089	132
	3	26	9	676	78
	1	21	1	441	21
	14	57	196	3249	798
	5	27	25	729	135
	4	23	16	529	92
	13	45	169	2025	585
	1	22	1	484	22
	11	43	121	1849	473
	8	35	64	1225	280
	1	19	1	361	19
	8	28	64	784	224
	12	44	144	1936	528
	13	40	169	1600	520
	3	35	9	1225	105
	9	48	81	2304	432
	2	25	4	625	50
	11	43	121	1849	473
	4	24	16	576	96
	5	30	25	900	150
	6	35	36	1225	210
	Σy =202	Σx= 981	Σy ² = 2168	Σx ² =35017	Σxy =7949

$$r = \frac{n \sum xy - \sum x \sum y}{\sqrt{(n \sum x^2 - (\sum x)^2) \times (n \sum y^2 - (\sum y)^2)}}$$

$$r = \frac{30 \times 7949 - 981 \times 202}{\sqrt{(30 \times 35017 - (981)^2) \times (30 \times 2168 - (202)^2)}}$$

$$r = \frac{238470 - 198162}{\sqrt{(1050510 - 962361) \times (65040 - 40804)}}$$

$$r = \frac{40308}{\sqrt{88149 \times 24236}} = \frac{40308}{46221} = 0.87207$$

As the value of $r = 0.87207$, the value of r^2 is 0.7605

Thus 76% of the variation in the length of service can be explained by the variation in age .

The SPSS printout looks like:

Correlations

		AGE	SERVICE
AGE	Pearson Correlation	1	0.872
	Sig. (1-tailed)	.	0
	N	30	30
SERVICE	Pearson Correlation	0.872	1
	Sig. (1-tailed)	0	.
	N	30	30

** Correlation is significant at the 0.01 level (1-tailed).

Thus $r = 0.872071$ and $r^2 = 0.760508$; therefore we have a strong positive correlation in which approximately 76% of the variation in length of service is explained by the variation in age. The remaining 24% will be due to other factors (e.g. not all employees will have 'worked their way up' through the company; some will have joined from other companies etc.). We can test the significance of this result by computing the test statistic and comparing it with critical values from the t-distribution. Our null hypothesis is that the true population correlation $\rho = 0$; our alternative hypothesis is that $\rho > 0$.

The test statistic $t = \frac{0.872071 \times \sqrt{(30-2)}}{\sqrt{(1-0.760508)}} = 9.4294$

Testing at a level of significance of 5%, the critical t-value is 1.701. Since the test statistic exceeds the critical value we can reject the null hypothesis, and conclude that we have evidence of a genuine linear relationship between length of service and age. Moreover, the fact that 9.4294 is much larger than 1.701 shows that we could adopt a much smaller level of significance and still reject the hypothesis of no relationship.

Note the output does not specifically provide the test statistic but does indicate that the correlation is statistically significant at a 1% level of significance.

Note we are using a one-tailed test of significance with alternative hypothesis $\rho > 0$; if r had been negative, our alternative hypothesis would be $\rho < 0$, the test statistic would also be negative, and the critical t-value would be on the negative side of the t-distribution. Thus the null hypothesis would be rejected if the test statistic is less than the critical value (i.e. closer to the tail). The test is therefore a complete mirror image of the example above.

Regression

Whilst correlation measures the extent to which there is a linear relationship between 2 variables, regression enable us to find the equation that describes that linear relationship.

The equation of a regression line has the form: $Y = a + bX$

where Y is the dependent variable (the one we wish to predict / explain) and X is the independent variable. The value " a " is known as the intercept of the line and " b " measures the gradient of this line.

The relevant formulae are:

gradient $b = \frac{\sum xy - \sum x \sum y}{n \sum x^2 - (\sum x)^2}$

intercept $a = \frac{\sum y - b \sum x}{n}$

Looking at the calculation above, we can see that the value of the "top" of b is the same as the "top" of r and that the value of the "left-bottom bracket" of r is the same as the "bottom" of b .

Thus

$$b = \frac{\sum xy - \sum x \sum y}{n \sum x^2 - (\sum x)^2} = \frac{40308}{88149} = 0.45727$$

$$\text{Thus } a = \frac{\sum y - b \sum x}{n} = \frac{202 - 0.45727 \times 981}{30} = -8.2194$$

Thus the equation of the line linking length of service (y) and age (x) is:

$$y = -8.2194 + 0.45727x$$

This equation can then be used to make predictions.

The SPSS regression output, with interpretation in italics, looks like:

Variables Entered/Removed^b

Model	Variables Entered	Variables Removed	Method
1	age ^a	.	Enter

Variables Entered/Removed

This simply tells us that 'age' was the independent variable and 'service' the dependent variable

Model Summary^b

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.872 ^a	.761	.752	2.629

Model summary

The value of the correlation coefficient, r was 0.872 and the value of r^2 was 0.761.

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	-8.219	1.657		-4.961	.000	-11.613	-4.826
	age	.457	.048	.872	9.429	.000	.358	.557

Coefficients

The unstandardized coefficients give us the values of a and b in the regression equation.

Thus the equation here is $y = -8.219 + 0.457x$

The final column "Sig" gives values less than 0.01 thus we can say that the coefficients of the regression equation are significantly different from zero at the 1% level (and thus at 5% level).

Casewise Diagnostics^a

Case Number	Std. Residual	service	Predicted Value	Residual
2	3.385	24	15.10	8.899

a. Dependent Variable: service

Casewise diagnostics

During the input dialogue, we asked for any standardised residuals outside the range -3 to + 3.

The output shows that one reading, case number 2, had a large standardised residual. This indicates that this point does not fit the general trend of the straight line and can be regarded as an outlier (i.e. an unusual reading). Case number 2 is an employee aged 51 with 24 years of service. On the scatter diagram, we can see that this point is a long way from the regression line.

Residuals Statistics^a

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	.47	17.85	6.73	4.603	30
Residual	-4.785	8.899	.000	2.583	30
Std. Predicted Value	-1.361	2.414	.000	1.000	30
Std. Residual	-1.820	3.385	.000	.983	30

a. Dependent Variable: service

Residual Statistics

This table can be ignored for simple cases.

Rank Correlation

If we wish to measure the strength of relationship between two variables, and at least one of them is ordinal, we need to use a nonparametric measure of correlation in which the strength of relationship is now based on the ranks of the data. There are two main measures of rank correlation, Kendall's τ statistic and Spearman's rank correlation coefficient; we shall focus on the latter.

Observed values for x and y are replaced by their ranks, and the difference (d) in ranking between each set of paired observations is calculated. Spearman's coefficient (r_s) is found from the following formula:

$$r_s = 1 - \frac{6 \sum d^2}{N(N^2 - 1)}$$

As with Pearson's coefficient, we can also test whether the result is statistically significant. Providing $N \geq 10$ we can use the t statistic as before:

$$t = \frac{r_s \sqrt{(N-2)}}{\sqrt{(1-r_s^2)}}$$

which has a t-distribution with N-2 degrees of freedom

For larger samples (say $N > 30$) we can use a normal approximation:

$$z = r_s \sqrt{(N-1)}$$

Example 1

[thanks to Richard Charlesworth for this example]

A survey of MBA students included a question which asked the respondent to identify the most important factor in their choice of UNL for their MBA.

Factor	Full-time	Part-time
MBA course programme/design	34	25
Discussion with tutor at recruitment fair/open evening	09	10
Recommended by a colleague/friend	22	18
Living in London	21	28
Credit transfer/flexible study	00	09
Prior study at UNL	04	10
Photos of the staff in the brochure	10	00

Factor:	Full-time MBA (%)	Part-time MBA (%)	FT ranks	PT ranks	d	d ²
MBA programme/design	34	25	1	2	-1	1
Discussion at fair/open eve	9	10	5	4.5	0.5	0.25
Recommendation..	22	18	2	3	-1	1
Living in London	21	28	3	1	2	4
CATS/flexible study	0	9	7	6	1	1
Prior study at UNL	4	10	6	4.5	1.5	2.25
Photos of the staff ..	10	0	4	7	-3	9
				10	$\Sigma d^2 = 18.50$	

$$r_s = 1 - \frac{6 \Sigma d^2}{N(N^2 - 1)}$$

Therefore $r_s = 1 - \frac{(6 \times 18.5)}{7 \times (49 - 1)} = 0.6696$

Note that tied ranks are averaged (i.e. the scores of 10% for Part-time MBA jointly cover the rankings 4 and 5, so both take on the rank 4.5).

SPSS Output: Spearman's rank correlation coefficient Nonparametric Correlations

Correlations

			FT MBA	PT MBA
Spearman's rho	FT MBA	Correlation Coefficient	1	0.667
		Sig. (1-tailed)	.	0.051
		N	7	7
	PT MBA	Correlation Coefficient	0.667	1
		Sig. (1-tailed)	0.051	.
		N	7	7