Early Human Development xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Early Human Development

journal homepage: www.elsevier.com/locate/earlhumdev

Best practice guidelines

WASP (write a scientific paper) using excel – 13: Correlation and regression

Victor Grech

Academic Department of Paediatrics, Mater Dei Hospital, Malta

ARTICLE INFO

Keywords:
Software
Computers
Statistics
Biostatistics

ABSTRACT

Correlation and regression measure the closeness of association between two continuous variables. This paper explains how to perform these tests in Microsoft Excel and their interpretation, as well as how to up these tests dynamically using Excel's functions.

1. Introduction

Correlation and regression are analysis techniques that measure the closeness of association between two continuous variables, but they are significantly different techniques.

Correlation tests for the association of a relationship between two variables and calculates an index describing the linear relationship between these variables as well as the strength of such a relationship. Regression predicts the value of a dependent variable based on the known value of an independent variable by calculating a line of best fit for the association between two variables.

2. Correlation

Correlation testing measures the strength and the direction of a putative linear relationship between two variables which may have independent units of measurement. It is important to note that correlation does not imply causation. The value of the degree of correlation (which is known as the correlation coefficient), ranges between −1 to 1. Either of the two extremes implies perfect correlation, perfect negative correlation and perfect positive correlation respectively.

The Pearson correlation test is most commonly used and this test assumes normally distributed data (for both variables). It is thus a parametric test. It is also known as 'linear or product-moment correlation' and the Pearson correlation coefficient is known as r. The test assumes a straight line relationship (rising or falling, respectively ranging from 1 to −1). This test can generally be used if there are more than 30 data points.

Some other correlation tests (all non-parametric with no assumptions with regard to the normality of the distribution of the data) include:

- Kendall's test with a correlation coefficient referred to as τ.
- Spearman's with a correlation coefficient referred to as $r_{s.}$

The following example illustrates Japanese data. Stress (including economic stress) has been shown to decrease male births due to an increased rate of male spontaneous abortions [1]. Japan's economy declined after the mid-1970s. The correlation between the economic decline (measured as annual change in percentage gross domestic product (GDP)) and the male to female ratio at birth (calculated as male divided by total births: M/F) is studied in Fig. 1 and the formulas used for the calculation are shown in Fig. 1 columns to the right [2]. There is clearly a correlation at statistically significant levels ($r^2 = 0.435$, $p = 0.0018$) but this does not in any way automatically imply causation.

Correlation is thus evidence for a possible causal relationship but correlation testing alone cannot indicate what that causal relationship (if any) might be. Indeed, causality is a subject that applies transversely across significant test results. Thus, all significant associations, not just correlations, could ultimately lead to causality discussions. Care must always be taken in the interpretation of statistically significant results, whichever test is used.

3. Linear regression

Linear regression provides the equation of straight line that best describes the association between two continuous variables. It models a relationship by providing a line of least squares fit for the available data. This tool has predictive value and provides a regression line and a regression equation ($y = a + bx$) (where a is the intercept and b is the slope). It also provides residuals i.e. the proportion of the change of the dependent variable that is actually dependent on the independent variable. It therefore also provides the proportion of the observed

E-mail address: victor.e.grech@gov.mt.

https://doi.org/10.1016/j.earlhumdev.2018.04.009

Please cite this article as: Grech, V., Early Human Development (2018), https://doi.org/10.1016/j.earlhumdev.2018.04.009

	A	B	C	D	E	F	G
1	Year	GDP growth/yr (%)	M/F		n	49	
2	1961	12.04	0.5144		r^2	0.435	=CORREL(B2:B50,C2:C50)
3	1962	8.91	0.5148				
4	1963	8.47	0.5137		df	47	=F1-2
5	1964	11.68	0.5143		t*	3.31	=F2*SQRT(F4/(1-F2^2))
6	1965	5.82	0.5129		2-tailed p	0.00180	=2*TDIST(ABS(F5),F4,1)
7	1966	10.64	0.5184				
8	1967	11.08	0.5129				
9	1968	12.88	0.5171				
10	1969	12.48	0.5173				
11	1970	-1.02	0.5172				
12	1971	4.70	0.5162				
13	1972	8.41	0.5157				
14	1973	8.03	0.5151				
15	1974	-1.23	0.5155				
16	1975	3.09	0.5149				
17	1976	3.97	0.5150				
18	1977	4.39	0.5147				
19	1978	5.27	0.5145				
20	1979	5.48	0.5150				
21	1980	2.82	0.5146				
22	1981	4.18	0.5143				
23	1982	3.38	0.5133				
24	1983	3.06	0.5138				
25	1984	4.46	0.5132				
26	1985	6.33	0.5136				
27	1986	2.83	0.5143				
28	1987	4.11	0.5141				
29	1988	7.15	0.5136				
30	1989	5.37	0.5137				
31	1990	5.57	0.5132				
32	1991	3.32	0.5139				
33	1992	0.82	0.5146				
34	1993	0.17	0.5136				
35	1994	0.86	0.5135				
36	1995	1.94	0.5126				
37	1996	2.61	0.5137				
38	1997	1.60	0.5126				
39	1998	-2.00	0.5132				
40	1999	-0.20	0.5135				
41	2000	2.26	0.5142				
42	2001	0.36	0.5133				
43	2002	0.29	0.5138				
44	2003	1.69	0.5133				
45	2004	2.36	0.5128				
46	2005	1.30	0.5130				
47	2006	1.69	0.5129				
48	2007	2.19	0.5137				
49	2008	-1.04	0.5128				
50	2009	-5.53	0.5131				

Fig. 1. Correlation calculation.

Table 1
Annual live births in Finland, 1953–1962.

Year	Births
1953	90,866
1954	89,845
1955	89,740
1956	88,896
1957	86,985
1958	81,148
1959	83,253
1960	82,129
1961	81,996
1962	81,454

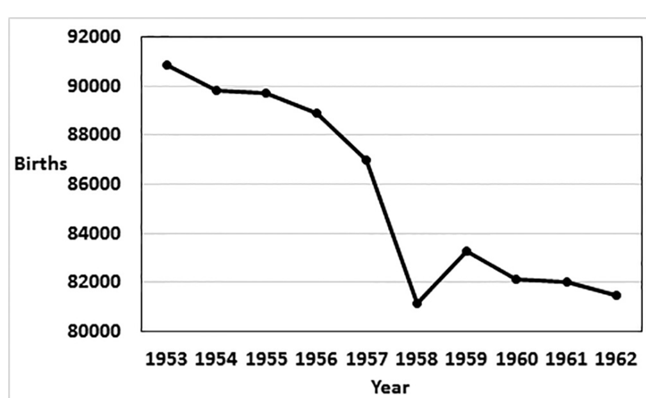

Fig. 2. Annual live births in Finland, 1953–1962 as per Table 1.