Correlation and Regression

Dale Easley University of Dubuque



Correlation does not imply causation

Towns with more doctors have more reported cases of disease.



Independent and dependent variables

рН	Germinated
6	0
6	3
6	4
6.5	1
6.5	2
6.5	4
7	2
7	4
7	6
7.5	5
7.5	6
7.5	7

Experiment: For a variety of pH values, count the number of plants germinating.

- 1. Design a lab experiment that gives the results shown.
 - Which is your independent variable? Why?
- 2. Design a field experiment that gives the results shown.
 - With what might your variables be autocorrelated?
- 3. Graph this with Excel.

Pearson's product moment correlation coefficient (r)

- 1. Paired data
- 2. Measure of correlation
- 3. Assumes data is normal about line fit
- 4. Impacted by number of data
- 5. Degrees of freedom = n-2



Try fitting a line to the data on the proceeding slide.

F looks at the ratio of variances, explained to unexplained



Is the r value significant?

_	А	В	С	D	E	F	G	Н	1	J					
1	pН	Num													
2	6	0													
3	6	3	Regre	ession			×								
4	6	4	Inpu	ıt					OK	1					
5	6.5	1	Ing	out <u>Y</u> Range:											
6	6.5	2	Inr	ut X Range:	t X Pange: EA41:4A412 E										
7	6.5	4		ar B Hanger		lauariau	415		Martin I.						
8	7	2	~	Labels		Constant is	<u>Z</u> ero		Helb						
9	7	4	~	Confidence Lev											
10	7	6	-												
11	7.5	5	Out	out options		(total		121							
12	7.5	6	(•	C New Worksheet Ply:											
13	7.5	7	C												
14			C	C New Workbook											
15			Re	siduals		_									
16			~	Residuals											
17				Standardized R	esiduals	l∾ Lin	e Fit Plots								
18			No	rmal Probability											
19			Normal Probability Plots												
20															
21			_							_					

Excel Output

SUMMARY	OUTPUT										~ 1 0	Pocidu			
Regression	Statistics										рпг	vesiuu		,	
Multiple R	0.689202								4	7					
R Square	0.475								<u>s</u> 2	-	:		٠	•	
Adjusted R	0.4225								p 0			•		•	
Standard E	1.630951								fesi	5,5	6	6,5	7	7.5	8
Observatic	12								· -2	1	•	•	•		
ANOVA									-4	_		P	н		
	df	SS	MS	F	ignificance	F			-						
Regressior	1	24.06667	24.06667	9.047619	0.013164										
Residual	10	26.6	2.66								pH L	ine Fit	t Plot		
Total	11	50.66667									P				
									5				•		
(Coefficients	andard Erro	t Stat	P-value	Lower 95%	Upper 95%	ower 95.09	pper 95.0%	Le -			1	•	MumSr	a clas
Intercept	-13.4333	5.704443	-2.35489	0.040303	-26.1436	-0.72304	-26.1436	-0.72304	Ĕ	:				• Hump	ALCK 3
pН	2.533333	0.842219	3.007926	0.013164	0.656752	4.409915	0.656752	4.409915	ž ²			•		Dradict	tod.
									0 -	E .	6.6		7.5	NumSc	ecies
									3.	.5	0.5		1.5		
									1			н			
RESIDUAL	OUTPUT				PROBABIL	TY OUTPU	г								
										No	rmal I	Proba	bility F	Plot	
bservation	cted Num(Residuals	dard Resid	uals	Percentile	NumGerm							, .		
1	1.766667	-1.76667	-1.13608		4.166667	0			e 8					•	
2	1.766667	1.233333	0.793115		12.5	1			E 6 -				. • • •	×	
3	1.766667	2.233333	1.436181		20.83333	2			Ĕ 4 -		•	• • •			
4	3.033333	-2.03333	-1.30757		29.16667	2			₹2-		• •				
5	3.033333	-1.03333	-0.6645		37.5	3			0 -	•			1		
6	3.033333	0.966667	0.62163		45.83333	4			0) 2	0 40	0 60	80	100	120
7	4.3	-2.3	-1.47905		54.16667	4					5	Sample Pe	rcentile		
8	4.3	-0.3	-0.19292		62.5	4									

Try another example

X=distance from road Y= traffic noise

х	У				
20	90				
40	90				
60	86				
80	81				
100	82				
120	75				
140	74				



Uncertainty in our prediction from a regression line

