



Correlation analysis with SPSS

Jarlath Quinn – Analytics Consultant

www.sv-europe.com



Just waiting for all attendees to join...



Jarlath Quinn – Analytics Consultant

www.sv-europe.com



FAQ's

- Is this session being recorded? Yes
- Can I get a copy of the slides? Yes, we'll email links to download materials after the session has ended.
- Can we arrange a re-run for colleagues? Yes, just ask us.
- How can I ask questions? All lines are muted so please use the Q&A panel if we run out of time we will follow up with you.











- Premier accredited partner to IBM, Predictive
 Solutions and DataRobot specialising in advanced analytics & big data technologies
- Work with open source technologies (R, Python, Spark etc.)
- Team each has 15 to 30 years of experience working in the advanced and predictive analytics industry

- Deep experience of applied advanced analytics applications across sectors
 - Retail
 - Gaming
 - Utilities
 - Insurance
 - Telecommunications
 - Media
 - FMCG



Agenda

- Why are correlation values useful?
- Interpreting correlation coefficients
- Estimating correlation values with bootstrapping techniques
- How correlations are calculated
- Automatically highlighting strong correlations
- Linear vs non-linear relationships
- Non-parametric correlations
- The limitations of correlations





Correlations

		Current Salary
Educational Level (years)	Pearson Correlation	.661**
	Sig. (2-tailed)	<.001
	N	474

**. Correlation is significant at the 0.01 level (2-tailed).

Why are correlations useful?

www.sv-europe.com

Why use correlations?

- Correlation is a term that we employ in everyday speech to denote things that *appear* to have some kind of relationship
- In analytics, correlations are specific values that are calculated in order quantify the relationships between variables
- This kind of analysis is powerful because, it allows us to <u>detect</u> and <u>measure</u> the strength of linear associations between an near infinite range of factors, such as:
 - Advertising spend and website hits
 - Product sales and competitor pricing
 - Vibration and component part failure
 - Rainfall and pollution
 - Study time and examination grade
 - Exercise and weight loss
 - Government spending and population health outcomes



The gateway to prediction

- Not only can we measure a linear relationship with correlation, but we can also use one variable to predict the other
- For example, if we know how much we're planning to increase our spend on advertising then we can use correlation to accurately predict what the increase in visitors to the website is likely to be.







Correlations

		Current Salary
Educational Level (years)	Pearson Correlation	.661**
	Sig. (2-tailed)	<.001
	N	474

**. Correlation is significant at the 0.01 level (2-tailed).

Interpreting correlations

www.sv-europe.com

Linear Correlation Scale



SN

Europe

510N

Pearson's r correlations



Pearson's r correlation coefficients



Non-Linear Relationships



-0.005

-.671

Pearson's r Correlations



A word of warning: always investigate the relationship





tile Er

- Analyze
 - Correlate
 - Bivariate



n	ployee with age.sav [\$Da	ataSet] - IBM SPS	S Statistics Data	a Editor						
	<u>E</u> dit <u>V</u> iew <u>D</u> ata	<u>T</u> ransform	<u>A</u> nalyze <u>G</u>	<u>G</u> raphs <u>C</u> ustom	<u>U</u> tilities	E	<u>x</u> tensions <u>W</u>	indow <u>H</u> elp		
Ì	🔚 🖨 🛄		Power A	nalysis		>	1	🛱 🕗 💽 [Q Search application	
			Meta An	alysis		>				
	🛷 id	💑 Q6	Descripti	ive Statistics		>	educ	⊿ iobcat	🛷 salarv	🖋 salbegi
	434	Male	Bayesiar	n Statistics		>	16	Clerica	\$34,950	\$20.2
	2	Male	Ta <u>b</u> les			>	16	Clerica	\$40,200	¢18.7
	2		Co <u>m</u> pare	Means and Propo	rtions	>	10	Olerica	4 \$40,200	φ10,7 C
	6	Male	<u>G</u> eneral Generali	Linear Model		>	15	Cierica	\$32,100	\$13,50
	8	Female	Mixed M	odels		>	12	Clerica	al \$21,900	\$9,7
	12	Male	<u>C</u> orrelate	9		>	+ Bivariate wi	ith Confidence Interva	als \$28,350	\$12,00
	13	Male	<u>R</u> egressi	ion -	-	>	10 Bivariate		\$27,750	\$14,2
	15	Male	Neural N	r etworks		>	🛃 Pa <u>r</u> tial		\$27,300	\$13,50
	16	Male	Classi <u>f</u> y			>	<u>6</u> Distances.	 Correlation	\$40,800	\$15.00
	17	Malo	<u>D</u> imensio	on Reduction		>		Clorico	\$46,000	¢14.00
	17	Male	Sc <u>a</u> le			>	15	CIETICA	al \$40,000	Φ14,23
	19	Male	Nonpara Eorecast	metric lests		>	12	Clerica	al \$42,300	\$14,2
	21	Female	<u>S</u> urvival	ung		>	16	Clerica	al \$38,850	\$15,00
	23	Female	M <u>u</u> ltiple	Response		>	15	Clerica	al \$24,000	\$11,10
	26	Male	Multiple	<u>V</u> alue Analysis			15	Clerica	al \$31,050	\$12,60
	28	Male	Complex	Samples		>	15	Clerica	al \$32,550	\$14,2
	30	Male	F Simulatio	on			15	Clerica	al \$31,200	\$14,2
	31	Male	Quality C	Control	ling	>	12	Clerica	al \$36,150	\$14,2
	33	Male	Direct M	ar <u>k</u> eting	g	>	15	Clerica	al \$42,000	\$15,00
	35	Male		08/2	2/61		17	Manage	er \$81,250	\$30,00
		F		00/0	7/00		•			

- Three variables chosen so three pairs of correlations
 - 1. Current Salary x Beginning Salary
 - 2. Current Salary x Education Level
 - 3. Beginning Salary x Education Level



	<u>V</u> ariables:	Ontions
 Employee Code [id] Date of Birth [bdate] Employment Category [jobcat] Months since Hire [jobtime] Previous Experience (months) [prevexp] Minority Classification [minority] Age of Employee [age] 	 Current Salary [salary] Beginning Salary [salbegin] Educational Level (years) [educ] 	Style <u>B</u> ootstrap Confidence interva
Correlation Coefficients		
Pearson Kendall's tau-b Spearman		

Correlations

		Current Salary	Beginning Salary	Educational Level (years)
Current Salary	Pearson Correlation	1	.880 ^{**}	.661**
	Sig. (2-tailed)		<.001	<.001
	N	474	474	474
Beginning Salary	Pearson Correlation	.880**	1	.633**
	Sig. (2-tailed)	<.001		<.001
	N	474	474	474
Educational Level (years)	Pearson Correlation	.661**	.633**	1
	Sig. (2-tailed)	<.001	<.001	
	N	474	474	474

**. Correlation is significant at the 0.01 level (2-tailed).



O a una lati a ma

	Corr	erations		minor mage
		Current Salary	Beginning Salary	Educational Level (years)
Current Salary	Pearson Correlation	1	.880**	.661 **
	Sig. (2-tailed)		<.001	<.001
	N	474	474	474
Beginning Salary	Pearson Correlation	.880**	1	.633 ^{**}
	Sig. (2-tailed)	<.001		<.001
	N	474	474	474
Educational Level (years)	Pearson Correlation	.661**	.633**	1
	Sig. (2-tailed)	<.001	<.001	
	N	474	474	474

**. Correlation is significant at the 0.01 level (2-tailed).



A SELECT INTERNATIONAL COMPANY

The table is a mirror image

Correlations

The diagonal values are all equal to one as they are the variables correlated against themselves

		Current Salary	Beginning Salary	Educational Level (years)
Current Salary	Pearson Correlation	1	.880 ^{**}	.661 ^{**}
	Sig. (2-tailed)		<.001	<.001
	N	474	474	474
Beginning Salary	Pearson Correlation	.880**	1	.633 ^{**}
	Sig. (2-tailed)	<.001		<.001
	N	474	474	474
Educational Level (years)	Pearson Correlation	.661**	.633 ^{**}	1
	Sig. (2-tailed)	<.001	<.001	
	N	474	474	474

**. Correlation is significant at the 0.01 level (2-tailed).



Correlations

The Significance values show how likely one is to get a correlation like that assuming there's no relationship between the variables

		Current Salary	Beginning Salary	Educational Level (years)
Current Salary	Pearson Correlation	1	.880 ^{**}	.661**
	Sig. (2-tailed)		<.001	<.001
	Ν	474	474	474
Beginning Salary	Pearson Correlation	.880**	1	.633**
	Sig. (2-tailed)	<.001		<.001
	Ν	474	474	474
Educational Level (years)	Pearson Correlation	.661**	.633 ^{**}	1
	Sig. (2-tailed)	<.001	<.001	
	N	474	474	474

**. Correlation is significant at the 0.01 level (2-tailed).



Th N values show how many cases the correlation was based on

Correlations

		Current Salary	Beginning Salary	Educational Level (years)
Current Salary	Pearson Correlation	1	.880 ^{**}	.661**
	Sig. (2-tailed)		<.001	<.001
	N	474	474	474
Beginning Salary	Pearson Correlation	.880**	1	.633**
	Sig. (2-tailed)	<.001		<.001
	N	474	474	474
Educational Level (years)	Pearson Correlation	.661**	.633**	1
	Sig. (2-tailed)	<.001	<.001	
	N	474	474	474

**. Correlation is significant at the 0.01 level (2-tailed).



- We can re-run the analysis, but this time....
 - Show only the bottom half of the matrix
 - Don't show the correlations of each variable against itself

🔚 Bivariate Correlations		Х
 Employee Code [id] Date of Birth [bdate] Employment Category [jobcat] Months since Hire [jobtime] Previous Experience (months) [prevexp] Minority Classification [minority] Age of Employee [age] 	Variables: ✓ Current Salary [salary] ✓ Beginning Salary [salbegin] ✓ Educational Level (years) [educ]	Options Style Bootstrap Confidence interval
Correlation Coefficients Correlation Coefficients Pearson Kendall's tau-b Spearman		
Test of Significance ● <u>T</u> wo-tailed ○ One-tai <u>l</u> ed		
✓ Elag significant correlations ✓ Show only the lo	ower triangle Show diagonal	



Correlations

		Current Salary	Beginning Salary
Beginning Salary	Pearson Correlation	.880 ^{**}	
	Sig. (2-tailed)	<.001	
	N	474	
Educational Level (years)	Pearson Correlation	.661 ***	.633**
	Sig. (2-tailed)	<.001	<.001
	N	474	474

**. Correlation is significant at the 0.01 level (2-tailed).



Let's explore Pearson's correlations in SPSS Statistics



$$r = rac{\sum \left(x_i - ar{x}
ight) \left(y_i - ar{y}
ight)}{\sqrt{\sum \left(x_i - ar{x}
ight)^2 \sum \left(y_i - ar{y}
ight)^2}}$$

How is a Pearson's correlation calculated?

www.sv-europe.com

Pearson's *r* (the most well known correlation measure)

- In statistics, the **Pearson correlation coefficient** is also known as **Pearson's r** or the **Pearson productmoment** correlation coefficient
- Correlations describe data moving together
- This is a **parametric** procedure. That means it makes assumptions about the data. Strictly speaking Pearson's *r* assumes
 - The level of measurement of the variables are continuous/scale (i.e. interval or ratio)
 - There should be no extreme outliers in the correlated variables
 - The data are <u>normally distributed</u> this is not needed for a reasonable sample size



Formula for Pearson's r

$$r = rac{\sum \left(x_i - ar{x}
ight) \left(y_i - ar{y}
ight)}{\sqrt{\sum \left(x_i - ar{x}
ight)^2 \sum \left(y_i - ar{y}
ight)^2}}$$

- r = correlation coefficient
- x_i = values of the x-variable in a sample
- $ar{x}$ = mean of the values of the x-variable
- y_i = values of the y-variable in a sample
- $ar{y}$ = mean of the values of the y-variable



Let's see an example of calculating Pearson's R





Correlations with rank order variables

www.sv-europe.com

Correlations for Ordinal Variables

- Rank order or 'ordinal' variables refer to variables such as rating scales
- These are not true numbers, but rather ranked 'numerals'
- Two techniques exist in SPSS that deal with these kind of data
 - Spearman's Rho
 - Kendall's Tau-b
- Both are non-parametric methods

🔚 Bivariate Correlations		×
Sex of respondent [sex] Year of survey [year]	<u>V</u> ariables:	<u>Options</u> Sty <u>l</u> e <u>Bootstrap</u> <u>C</u> onfidence interval
Correlation Coefficients		
● <u>T</u> wo-tailed ○ One-tai <u>l</u> ed		
Elag significant correlations Show only the lower OK Pa	er triangle 🗹 Show diagonal ste <u>R</u> eset Cancel Help	



Spearman's Rho

- Spearman's Rho works by ranking the original values from the lowest number to the highest
- For this reason, it's sometimes referred to as Spearman's Rank Correlation
- Unlike Pearson's Correlation here we don't know if the variables are linearly related as ranking the values hides this relationship
- Spearman's correlation detects *monotonic* relationships. A monotonic relationship is one in which, as the size of one variable increases, the other variables also increases, or where the as the size of one variable increases, as the other variable decreases.
- Spearman correlations are not affected by outliers but analysts should still consider whether extreme outliers are valid reflections of the population under consideration



Spearman's Rho

• Consider this non-linear relationship....





Let's explore how Spearman's correlations work

Kendall's Tau

- **Kendall's Tau b** also works by ranking the original values from the lowest number to the highest
- However, this time the analysis focuses on *the degree of concordance and discordance* between two ranked columns of data

	Movie_Title	IMDb	Rotten_Tomatoes
1	Uncorked	1	
2	Spenser Confidential	2	R >1
3	The Willoughbys	3	
4	Tigertail	4	→3
5	Extraction	5	2
6	The Half of It	6	8
7	To All the Boys I've Loved Before	7	r -9
8	LA Originals	8	4
9	Miss Americana	9	7
10	Crip Camp: A Disability Revolution	10	← → 10



Kendall's Tau

- Some argue that the significance estimates and confidence intervals for Kendall's Tau tend to be more reliable than for Spearman correlations.
- Kendall's Tau tend to give smaller correlation values than Spearman's and can also take much longer to calculate

	A Movie_Title	IMDb	Rotten	Tomatoes
1	Uncorked	1	*	6
2	Spenser Confidential	2	R	>1
3	The Willoughbys	3	\sim	75
4	Tigertail	4	R	→3
5	Extraction	5		2
6	The Half of It	6		8
7	To All the Boys I've Loved Before	7	K	79
8	LA Originals	8		4
9	Miss Americana	9	K	7
10	Crip Camp: A Disability Revolution	10	<	→ 10



Let's compare Kendall's Tau to Spearman's correlations

Visualising Correlations for Ordinal Variables





No matter how fast our income goes up we never seem to get ahead

May require a different approach than scatterplots





The Limitations of Correlations

www.sv-europe.com

Correlation does not indicate causation



https://www.productleadership.com/does-causation-imply-correlation/



Can't accurately measure curvilinear relationships





Are influenced by the range of values in the sample



Educational Level (years) by Beginning Salary



Are influenced by the range of values in the sample



Educational Level (years) by Beginning Salary



Can be unduly affected by extreme/outlier values



Current Salary by Beginning Salary



Can be unduly affected by extreme/outlier values



Current Salary by Beginning Salary



Working with Smart Vision Europe Ltd.

• Sourcing Software

- You can buy your analytical software from us often with discounts
- Assist with selection, pilot, implementation & support of analytical tools
- <u>http://www.sv-europe.com/buy-spss-online/</u>
- Training and Consulting Services
 - Guided consulting & training to develop in house skills
 - Delivery of classroom training courses / side by side training support
 - Identification & recruitment of analytical skills into your organisation
- Advice and Support
 - offer 'no strings attached' technical and business advice relating to analytical activities
 - Technical support services





Contact us:

+44 (0)207 786 3568 info@sv-europe.com Twitter: @sveurope Follow us on Linked In Sign up for our Newsletter

Thank you

www.sv-europe.com