Research Methods II

Correlation to Regression

Stacey Coleman

Overview

- Recap of correlation analysis
- Introduction to regression

Correlation

- Non-experimental method
- With correlational studies, no IV is manipulated
  - There are no IV's or DV's
  - Only **co-occurring** variables
  - Requires data at **interval** level

Variables in correlation/regression designs

- In correlation and MR, we tend not to talk about IV's and DV's…
  - Co-related or co-existing variables
  - Predictor variables (IV)
  - Criterion variable / Outcome variable (DV)

Correlational research

Three components to correlation research:

1. Researcher measures one variable (X)
2. Researcher measures a second variable (Y)
3. Researcher statistically determines whether (X) and (Y) are related

Correlation

- Important!
  - Correlation does **NOT** indicate ***causation*** – we cannot say something has caused something else!

- There may be a third variable more suited to explain the relationship between the variables of interest

Correlation example

Is there a relationship between levels of self-esteem and depression?

Correlation example

Is there a relationship between levels of self-esteem and depression?

Not manipulating anything, simply measuring levels of each variable

Self-esteem (X) ⟺ ***Level of depression (Y)***

✓

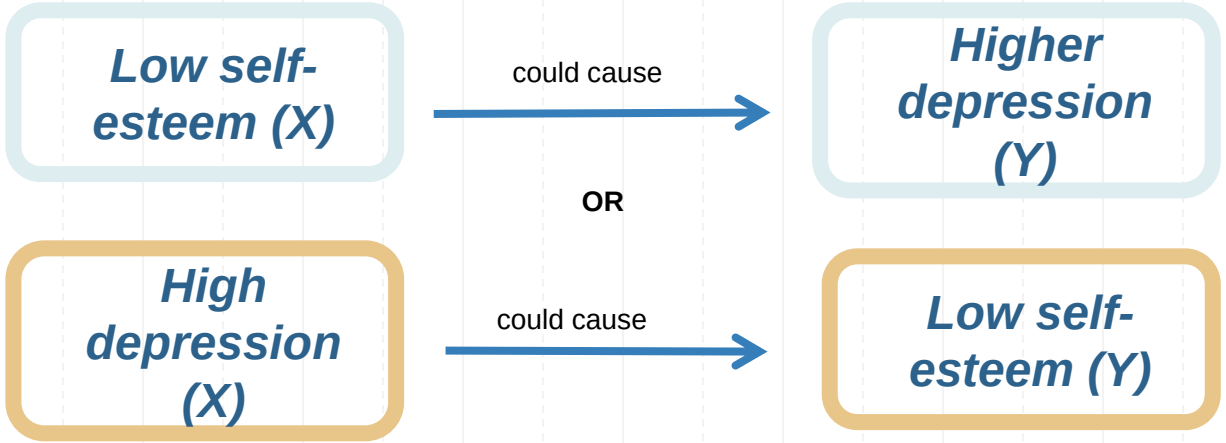
Correlation example

- So what can we say about these two variables?

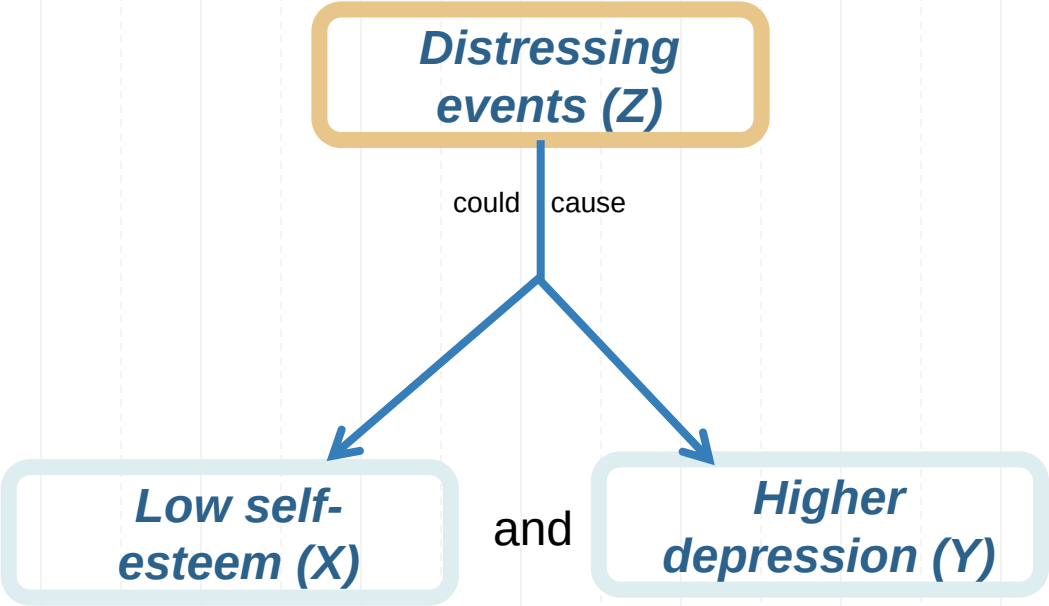
We can say that levels of self-esteem and depression are related

Correlation example

- What **can't** we say?
- We are not exploring whether one **causes** the other, so we cannot determine causality.



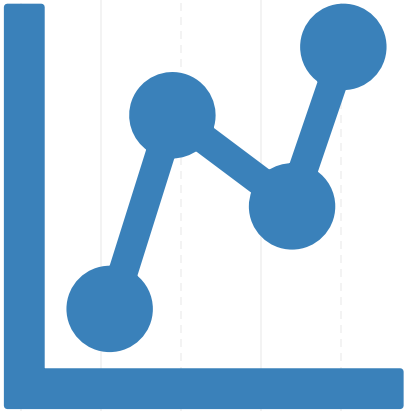
OR!


Distressing events (Z)
could cause
Low self-esteem (X)
and
Higher depression (Y)

Refresher

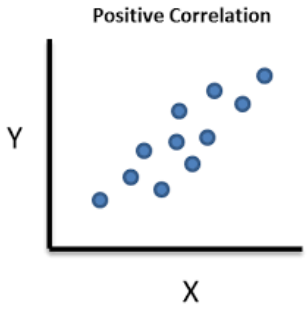
- Nature of correlational designs:
  - Is there an association/relationship?
  - What is the ‘direction’ of the relationship?
  - What is the ‘strength’ of the relationship?
  - Is it statistically significant?

The Scatterplot

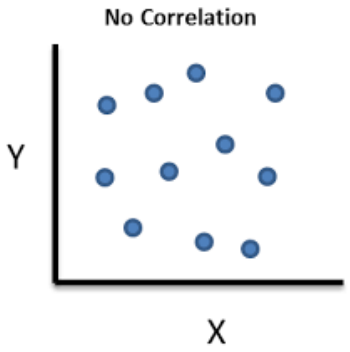


- An essential part of any preliminary analysis when conducting correlational research.
- Scatterplot = graph that plots each participants score on one variable against their score on the other variable.
- The Scatterplot allows you to see whether:
  - There is a relationship between 2 variables
  - What that relationship is
  - What the spread of the data is like (are there outliers)

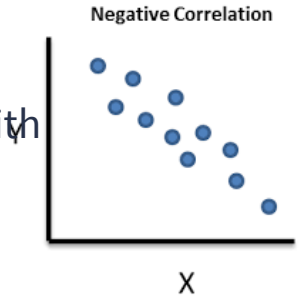
Types of correlation/relationship



Higher scores on one variable are associated with higher scores on a second variable



Higher scores on one variable are associated with lower scores on a second variable



The two sets of scores are completely unrelated

Correlation Coefficient

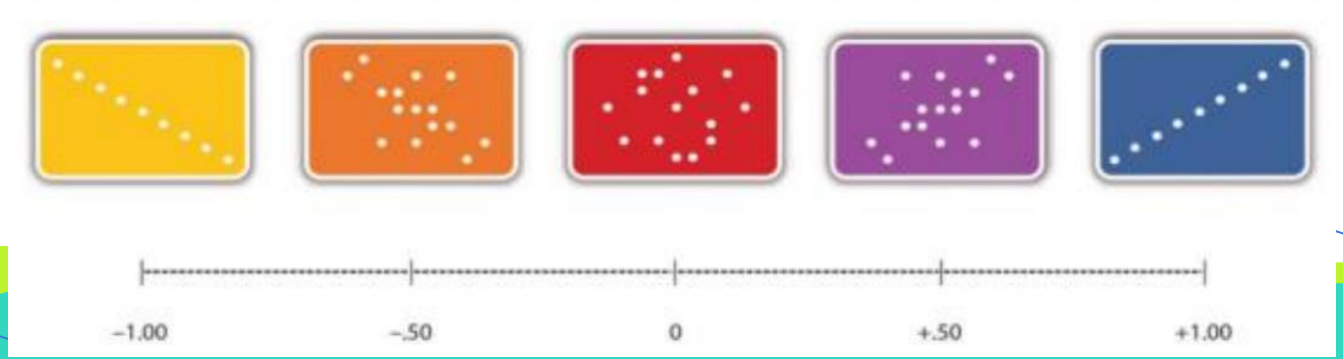
Perfect correlation = 1

Sign simply denotes direction; negative or positive

Perfect	Strong	Moderate	Weak	No relationship	Weak	Moderate	Strong	Perfect
-1	-0.8	-0.6	-0.4	-0.2	0	0.2	0.4	0.6	0.8	1

Increasing strength

Increasing strength



Correlation

- The **strength** of the relationship between two variables is the degree to which one variable does tend to vary with the other
  - If one increases, does the other increase/decrease, too?

- Strength is expressed on a scale ranging from -1 (perfect negative correlation) to +1 (perfect positive correlation)
  - The sign (+/-) indicates direction
  - 0 means there is no correlation
  - The numerical value indicates the strength

- This final number is referred to as a ***correlation coefficient*** and is represented as ***r***

Correlation Coefficient

- You will always get a value between -1 and 1
- This value is just a descriptive value
- Guide to describing r-values (Cohen, 1988)
    - r = ± .10 to .29 – small / weak
    - r = ± .30 to .49 – medium / moderate
    - r = ± .50 to 1.00 – large / strong

Data associations: From correlation to regression

What is regression?

- Regression is used to discover the effect of one variable (X) on another variable (Y)
- So, similar to correlation: **However!**
  - Correlation allows use to determine how strongly 2 variables are related
  - Linear regression also tells us **by HOW MUCH** one variable will change if the other variable changes
  - Aim is to ***predict*** scores on one variable based on scores from another variable

What is regression?

- So – if *x* changes by a certain amount, we will be able to estimate the amount in which *y* will change
- Regression can also be used to analyse a more ‘causal’ relation between variables
- Theory suggests a DV and an IV

Regression example

- Imagine you are working as a marketing psychologist and your job is to look at whether the price of your company's soft drink relates to the amount of cans that are bought

Regression example

- You would imagine that as the price increased then sales would decrease
- Correlation would be able to show us this
  - i.e., negative correlation
- Linear regression can show us the **amount** of the change

Regression example

- E.g. if the price increases by 50%, sales will fall by 40%
  - Allows us to predict *y* from *x*
- If we know by how much sales decrease, as a result of every penny increase, then we can predict sales from price
  - E.g., for every 2p increase, sales decreased by 100 cans

Regression example

- Conducting a linear regression analysis to examine this would allow us to suggest that a score on one variable influenced the score on the other variable

- In this way, we can try to infer causal relationships (but only if the variables have been manipulated..)

Linear Regression

- Proceed in the same way as with a correlation:
  - Two variables are examined to see if there is a linear relationship between them

IMPORTANT TERMINOLOGY!

Criterion Variable

- The variable you want to predict
- Equivalent to the DV
- Also called *y*

Predictor Variable

- The variable that is predicting the DV
- Equivalent to the IV
- Also called *x*

- **ONE predictor variable and ONE criterion variable = bivariate (linear) regression**
  - **Two or more predictor variables = multiple regression**

IMPORTANT TERMINOLOGY!

b

- The amount by which *y* changes, when *x* changes by 1 unit
- Represented by the slope/tilt of the straight line in a scatterplot
- Also known as the regression coefficient (regression weight)
- The larger the value of b, the steeper the slope
- The steeper the slope the more y changes as a result of x
- Using a straight line allows us to predict a score on *y* from *x* (SPSS will draw the line in the right place)

IMPORTANT TERMINOLOGY!

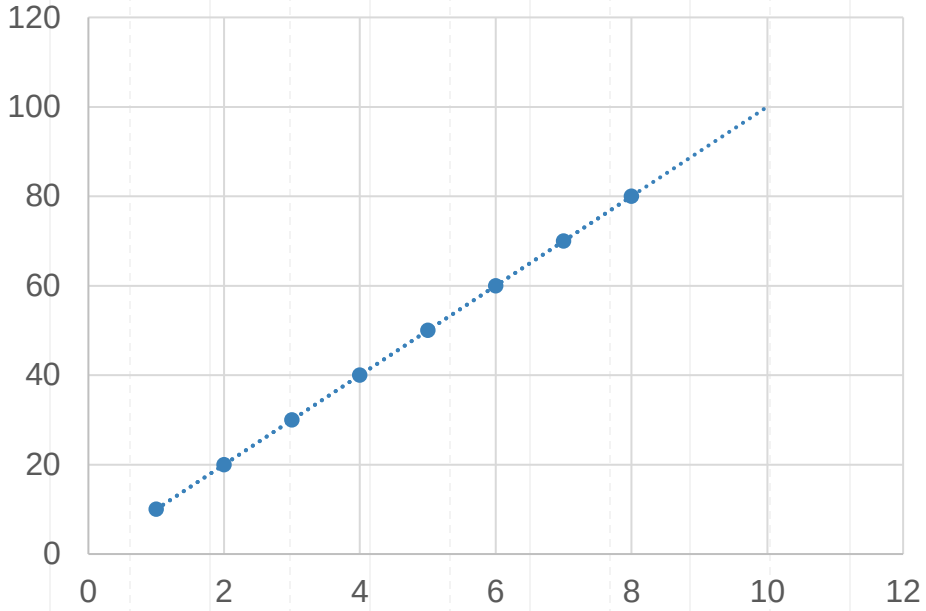
a

- The intercept (also called the Constant)
- This is the value of *Y'* when X is 0
  - The symbol *Y'* (pronounced y prime) is used in an equation when we want to predict values of *y* from our observed values of *x*
- When regression line is added to scatterplot, ***a*** is the value of *y* at the point where the line intercepts at the Y-axis
- NB: by default, SPSS applies axis scales around values in the dataset so the scatterplot axis may not include a 0. In such cases, *a* cannot be read from the chart – but SPSS provides the value for *a* elsewhere.

Perfect relationship - Positive

- A perfect positive correlation is when it is possible for the straight line to pass through every data point
  - E.g., if every time *x* increased by 1, *y* increased by 10.

Perfect Relationship - Positive



The point where the line intersects the y axis is called *a*. Here *a* = 0

The Regression Equation

- From linear regression analysis, we obtain a regression equation
- This shows how *y* changes as a result of a change in *x*.
- From the formula we can then calculate someone's score on *y* from their score on *x*.

The Regression Equation

- Using the below formula, we can calculate a score on *y* from a score on *x:*

$$y = a + (b*x)$$ **[positive]**

$$y = a - (b*x)$$ **[negative]**

The Regression Equation

- Example:
- Say the slope of the line is 10 (so, $b = 10$)
- And $a = 0$ (starting point on the chart, intercept, constant)
- A person scores 7 on our IV (x)

The Regression Equation

- $b = 10$
- $a = 0$
- $x = 7$

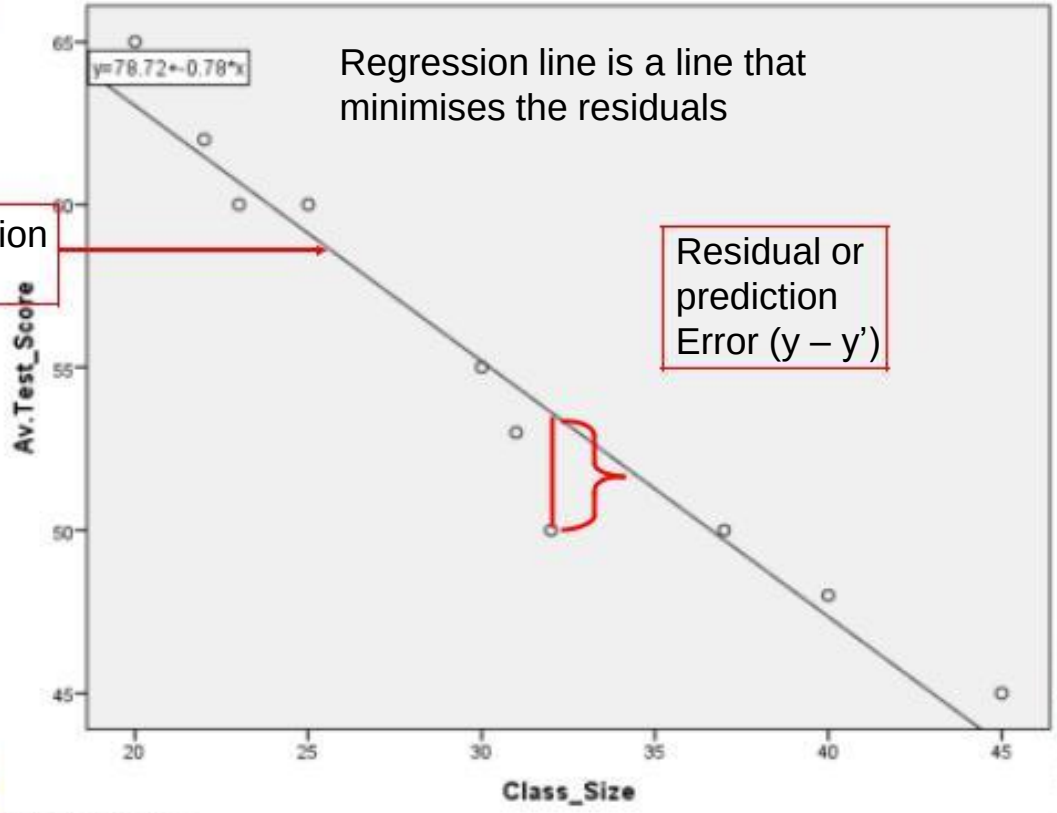
- ***y = a + (b*x)***

- **y = 0 + (10*7) = 70**

Of course, it's easy to predict a score when all of the points fall on a straight line…

Residuals (Y – Y’)

- Unless we have a perfect relationship, then we will always make some errors when using the line of best fit to predict a score on *y*
- Data points that are off the line represent errors of prediction
- Using the line of best fit to predict values will give us a predicted value
- The difference between the actual scores (Y) and predicted scores (Y’) are called **residuals**.
  - Can be positive or negative



Regression line is a line that minimises the residuals

Regression line

Residual or prediction Error (y – y')

UNIVERSITY OF WALES
Trinity Saint David

Non-perfect relationships

- Easy to predict from perfect relationships; but if your points don't all fit exactly on the line then…
  - SPSS will still insert a line of best fit
    - Gives the best fit through all data points
  - Calculated by minimising the vertical distances from the data points and the line
  - Called the **least-squared regression line**

The SPSS Output

Important Parts of the SPSS Output

Model Summary

- Provides the correlation between *x* and *y*
  - *Shows how well the data points cluster around the line of best fit*
  - *Presented as a simple Pearson's r*

- Provides the square of the correlation coefficient (R squared)
  - This gives you the ***variance explained***

Important Parts of the SPSS Output

Model Summary

- Also give the **Adjusted R Square** value
  - Accounts for no* of participants and variables in the analysis
- We use and report the Adjusted R square value – why?
  - R squared is too optimistic, as the line of best fit is based on a sample not the whole population
  - We want the results to generalise to the population, so the adjusted R square adjusts the figure to give a more realistic estimate

Model Summary

- We use and report the usted R square value
- ***Why?***
- R squared is too op
  on a sample, not
- We want the res
  the adjusted R s
  realistic estimate

What percentage of the variance is explained by the model as a whole?

Convert adjusted R squared to a %

Standard Error

- Remember, the stats are not free from error.
  - Results are analysed with a particular sample and it may be that a sample taken on a different day would give slightly different results.
  - With repeated measures we would find a range of values, the standard deviations of such distributions are called the standard error.
  - The Standard error gives us a measure of how accurate our estimation is likely to be

ANOVA table

- ANOVA summary table shows whether the regression line is significantly different from 0
  - i.e., it predicts better than would be expected by chance
- If the slope $b = 0$, then the line of best fit is horizontal
- If the value is significant then it is unlikely the results would have arisen by chance or sampling error

Coefficients box

Values for *a* and *b*

- The value of *a* is given in the ‘Constant’ row (under ‘unstandardized coefficients B’ column
  - When the modelling process is applied to the values of the dataset, the criterion value for someone who scores 0 on the predictor value would be this value in the a box.
- The value for *b* shows that for every 1 unit increase, *b* will either increase or decrease (depending on the sign) by that value
- The slope, *b*, has also been converted to a standardised Beta score, which allows us to compare the data in terms of SD

Coefficients box

Confidence limits

- Based on the standard error
- Confidence limits tell you that your population slope can be found within the range of 2 values

- You can now work out/predict new cases using the *a* and *b* values from your output and entering them into the regression equation

$$y = a + (b^{*}x)$$