

Linear Regression

Dr. Kate Williams

Regression

- Used to discover the effect of one variable (x) on another variable (y).
- Similar to correlation: **However!**
 - Correlation only allows us to conclude how strongly two variables relate to each other (in terms of magnitude and direction)
 - Linear Regression also tells us **by HOW MUCH** one variable will change if the other variable changes.
 - Aims to **predict** scores on one variable based on scores from another variable.

Regression

- Basically – if x changes by a certain amount we will be able to estimate the amount in which y will change.
- Imagine you are 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 being bought.



Regression

- You would suspect that as the price increases then sales would decrease.
- A simple correlation would be able to show us this – i.e., a negative correlation.
- What linear regression does is show us the amount of the change.
 - E.g., if the price of the drink increases by 50%, sales will fall by 40%.
 - Allows us to predict y from x

Regression

- 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 in price, sales decrease by 100 cans.
- 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.
- We can try to infer causal relationships (but only if the variables have been manipulated).



Regression

- Proceed in same way as correlation:
 - i.e., Two variables examined to see if there is a linear relationship.
- The variable you want to predict is called the **Criterion** variable (Dependent variable)
 - Also called y .
- The variable that is predicting the DV is called the **Predictor** variable (or explanatory) variable.
 - Also called x
- **ONE predictor variable and ONE criterion variable is a bivariate regression.**
 - (Two or more predictor variables = multiple Regression)

Regression

- The predictor variables can be measured using a range of scales
 - Although ideally at interval or ratio level
- However, the criterion variable must be measured using an interval or ratio scale.
- As with a correlation, regression does not imply a causal relationship
 - Unless variables have been manipulated.

Important terminology!

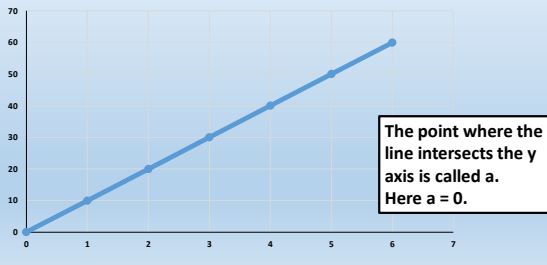
- **b** = The amount by which y changes, when x changes by 1 unit
 - Represented by slope/tilt of straight line in scatterplot
 - Also known as regression coefficient (regression weight)
 - Larger value of b = steeper the slope.
 - Steeper slope = more y changes as result of x .
 - Using straight line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

Important Terminology!

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

Perfect Relationships - Positive

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



The Regression Equation

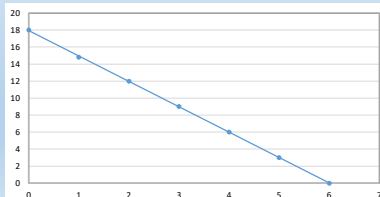
- From linear regression analysis, we obtain a regression equation
 - This shows how y changes as a result of a change in x .
- From the formula we can then calculate someone's score on y from their score on x .
- The general formula is:
 - $y = (b \cdot x) + a$ or $y = a + (b \cdot x)$
- Say the slope of a line is 10.
 - i.e., every time x increases by 1 unit, y changes by 10.
- The equation shows y can be predicted by multiplying a score on x by b and then adding the constant, which is a .
- So, in this example:
 - $a = 0$ (starting point on the chart, intercept, constant)
 - $b = 10$
 - $y = a + 10 \cdot x$

The Regression Equation

- So, if we know that a person has a score of 7 on x , then we are able to predict their score on y :
 - $y = a + (b \cdot x)$
 - $y = 0 + (10 \cdot 7) = 70$
- We can also predict negative relationships:

$$y = a - (b \cdot x)$$

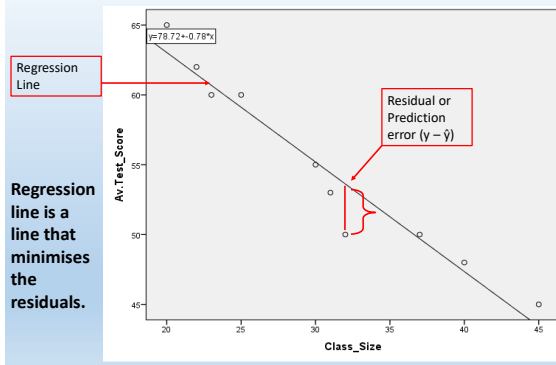
The minus sign
= a negative
relationship



Residuals ($Y - Y'$)

- If we use our line of best fit to predict a score, unless we have a perfect relationship we will always make errors.
- Data points off the line represent errors of prediction.
- If we use the line of best fit to predict sales for each price of soft drink for example, then we will obtain a predicted sales figure.
- The difference between the actual scores (Y) and predicted scores (Y') are called residuals.
- They can be positive or negative.

Residuals



Non-perfect relationships

- Easy to predict from a perfect relationship
 - But human behaviour is inherently noisy.
- So, if all points do not fit exactly on the line then what can you do?
- SPSS can still draw a line of best fit – i.e., a line that gives the best fit through all the data points.
- SPSS calculates this by minimising the vertical distances from the data points and the line.
- This is called the least-squared regression line.

Important parts of the SPSS output!

•Model Summary box

- Provides 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 (in the R squared column).
 - This gives you the *variance explained*.
- Also gives you the Adjusted R square value

Important parts of the SPSS output!

•Model Summary box

- Adjusted R square
 - Accounts for the number of participants and variables in the analysis.
- 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.

Standard Error

- Remember, stats are not free from error.
 - Results are analysed with a particular sample and maybe a sample taken on a different day would give 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.

Important parts of the SPSS output!

•ANOVA table

- The 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 sampling error.

Important parts of the SPSS output!

•Coefficients box

- Provides the values for a and b .
 - The value for b shows that for every 1 unit increase, b will either increase or decrease (depending on whether the value is negative or not) by that value.
- Also provides confidence limits (based on the standard error)
 - Confidence limits tell you that your population slope can be found within the range of 2 values.
- The slope, b , has also been converted to a standardised Beta score, which allows us to compare the data in terms of a Standard Deviation

Important parts of the SPSS output!

•Coefficients box

- The value of a is given in the 'Constant' row under the '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.
- You can now work out/predict new cases using the a and b values from your output and entering them in to the regression equation.
 - $y = (b \cdot x) + a$
