SO5041 Unit 15: Correlation

Brendan Halpin, Sociology

Autumn 2019/0

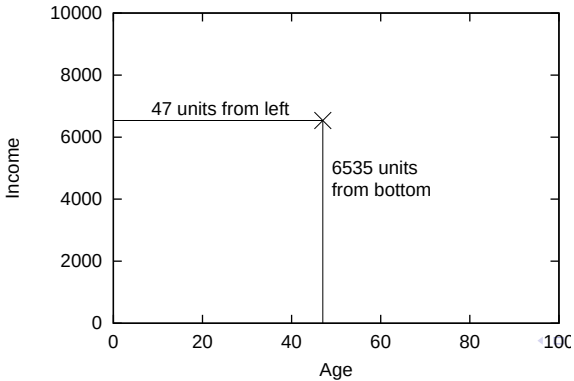
Remaining topics

- From today we look at two related techniques for interval and ratio variables:
  - Correlation: a single measure of association for interval/ratio variables
  - Linear Regression: a very powerful technique for describing the relationship between one interval/ratio variable and another

Scatterplots

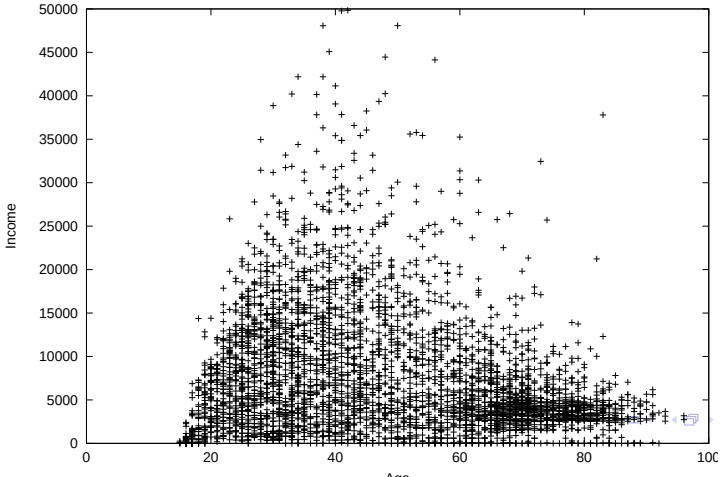
- Scatterplots can be considered as interval/ratio analogue of cross-tabs: arbitrarily many values mapped out in 2-dimensions

Figure 1: Age 47, income €6,535.



Age vs Income

Figure 2: Age and income, Wave 1 BHPS.

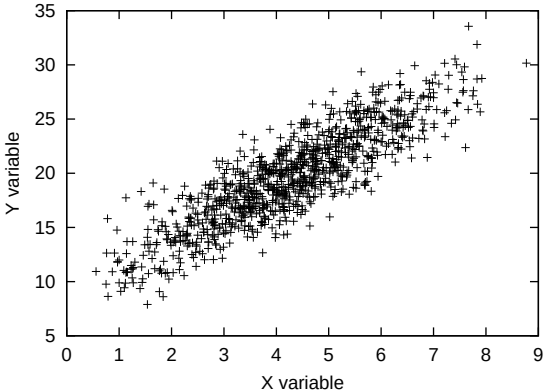


Summarising association simply

- We can see a lot of detail in a scatterplot, but sometimes we can summarise it in simple ways
- For instance the two variables may have a positive association: when one is high the other tends to be high, and vice versa
- Or a negative association: when one is high the other tends to be low

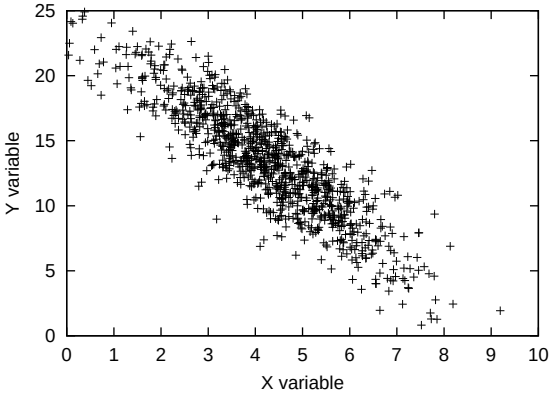
Strong positive

Figure 3: Fictional data displaying a strong *positive* linear relationship.



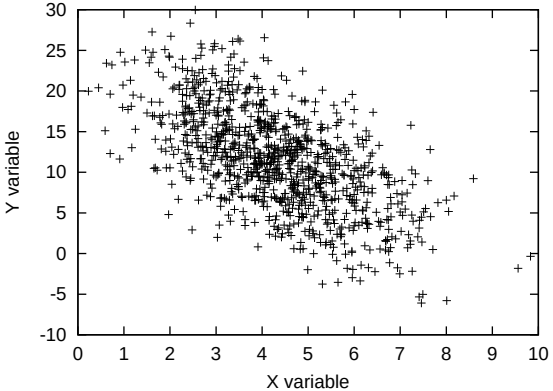
Strong negative

Figure 4: Fictional data displaying a strong *negative* linear relationship.



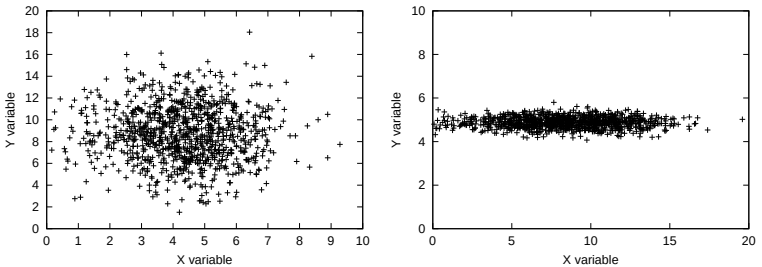
Weak negative

Figure 5: Fictional data displaying a *weak* negative linear relationship.



No relationship

Figure 6: Fictional data displaying absence of a relationship.



App

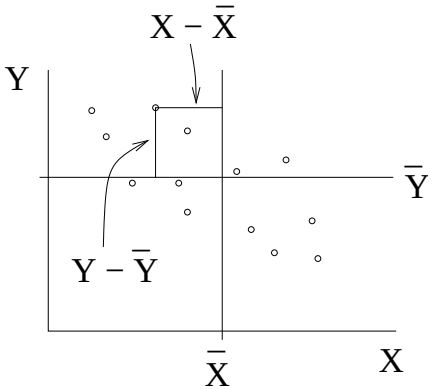
http://teaching.sociology.ul.ie:3838/so5041/corr

The correlation coefficient

- How does it work?
- Combines X deviations $(X_i - \bar{X})$ and Y deviations $(Y_i - \bar{Y})$ – i.e., compares each point with the mean for X and the mean for Y
- With positive association, cases below average on X tend to be below average on Y (and above average on X tend to be above average on Y)
- With negative association, cases below average on X tend to be above average on Y and vice versa
- With positive association below the mean, both $X_i - \bar{X}$ and $Y_i - \bar{Y}$ are negative, so $(X_i - \bar{X})(Y_i - \bar{Y})$ is positive
- With negative association, $X_i - \bar{X}$ and $Y_i - \bar{Y}$ tend to have opposite signs, so $(X_i - \bar{X})(Y_i - \bar{Y})$ is negative

Deviations

Figure 7: X and Y deviations for correlation



Pearson Product-Moment Correlation

Coefficient

- Pearson Product-Moment Correlation Coefficient (r)

$$r = \frac{SXY}{\sqrt{SXX.SYY}}$$

$$SXX = \Sigma(X - \bar{X})^2$$

$$SYY = \Sigma(Y - \bar{Y})^2$$

$$SXY = \Sigma(X - \bar{X})(Y - \bar{Y})$$

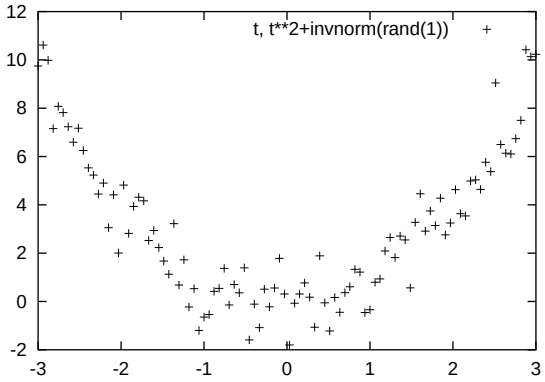
- Range: $ -1 \leq r \leq +1 $
- r is a symmetric measure: $r_{xy} = r_{yx}$

Pitfalls

- Correlation is not causality
- Absence of correlation is not absence of relationship: non-linearity

Non-linear!

*Figure 8: * Fictional data displaying a strong *non-linear* relationship and a near-zero correlation.



Estimating correlations in Stata

```
. pwcorr ofimn ojbhgs ojbhrs, sig
```

	ofimn	ojbhgs	ojbhrs
ofimn	1.0000		
ojbhgs	0.4851 0.0000	1.0000	
ojbhrs	0.4245 0.0000	0.2489 0.0000	1.0000

Viewing the same correlations