Bi-variate analysis and linear regression

- Statement for Audio and Video Learning Resources
- Video and audio content at the University uses closed captions generated by automatic speech recognition (ASR). The ASR process is based on machine learning algorithms which automatically transcribe voice to text. According to our technology providers, this process is approximately 70-90% accurate depending on the quality of the audio, and consequently video and audio closed captions may include some transcription errors. It is therefore important to recognise that the original recording is the most accurate reflection of the content, and not the captions.
- If you require accurate captions as part of your reasonable adjustments, please contact the Inclusion Centre to discuss your requirements.

Bivariate Data and Linear Regression

Ref: D. Diez, M. Cetinkaya-Rundel, C. Barr, *OpenIntro Statistics* (4th Edition), OpenIntro, 2019 [available for download at www.openintro.org/book/os/, accessed 27/01/2020]

Content

- Bivariate Data
- Measuring Associations between variables
- Linear Regression
  - Finding the 'best straight line'
- Visualisation
- Residuals
- Generalisations

Bivariate Data

- Bivariate – involving 2 variables
  - Data consists of pairs of values (x,y)
  - Where x is a value of some variable X and y is a corresponding value of some variable Y
- Examples
  - People’s size: height and weight
  - Pressure and temperature measured at a number of locations
  - GPS location: latitude and longitude
  - Game result: goals for team 1 and goals for team 2.

Goals

- Individual variables X and Y can be analysed independently using univariate methods
  - But independent analysis of each variable does not make use of the information about pairs
- The relationship between the 2 variables can be discovered, visualised and quantified

Measures of Association

- Covariance

$$\sigma(x, y) = \frac{\sum_{i=1}^{N}(x_i - \bar{x}) * (y_i - \bar{y})}{N}$$

- Measure of how much two variables vary together.

Covariance interpretation

- **Positive:** there is a positive association between variables x and y.
  - x tends to be higher than average when y is higher than average
- **Negative:** there is a negative association.
  - x tends to be lower than average when y is higher than average
- **0:** no association between x and y

Notes on Covariance

- The covariance is here averaged over N data points
- Can also define covariance as the expected value for a theoretical distribution
  - Data can be treated as sample (cf. St. Dev.)
- Related to variance by $\boldsymbol{\sigma(x, x) = \sigma^2(x)}$
  - Variance is square of standard deviation

Correlation

- Not obvious what size of covariance means
  - When is it close to zero?
- Can normalise by dividing by standard deviation of the variables
  - This is the correlation (coefficient)
  - $R(x, y) = \frac{\sigma(x,y)}{\sigma(x) * \sigma(y)}$
  - $-1 \leq R \leq 1$

Interpretation of Correlation

- **1:** perfect linear relationship with positive gradient
- **-1:** perfect linear relationship with negative gradient
- **0:** no (linear) relationship
- R is sometimes called the **product-moment correlation coefficient**

Linear Regression

- It is often useful to fit some function $\mathrm{y}(\mathrm{x})$ to the data
- In regression, we assume some general functional form and fit parameters to the data
- For example, **linear regression**
  - Assume form is $\boldsymbol{y} = \boldsymbol{\beta_0} + \boldsymbol{\beta_1}\,\boldsymbol{x}$
  - Use the data to determine best values of coefficients $\boldsymbol{\beta_0}$ and $\boldsymbol{\beta_1}$

Terminology

- X is the predictor (variable) and Y is the response (variable)
- We are regressing Y on X
  - NOT X on Y
- If we swap the roles of Y and X, we get a different line with
  - Y as predictor and X as response
  - Regression of X on Y

Assumptions

- y obeys the model

$$y = \beta_0 + \beta_1 x + \varepsilon$$

Where ε is a random variable and individual ε_i are

- Independently
- Identically
- Distributed according to a Normal distribution
  - with zero mean

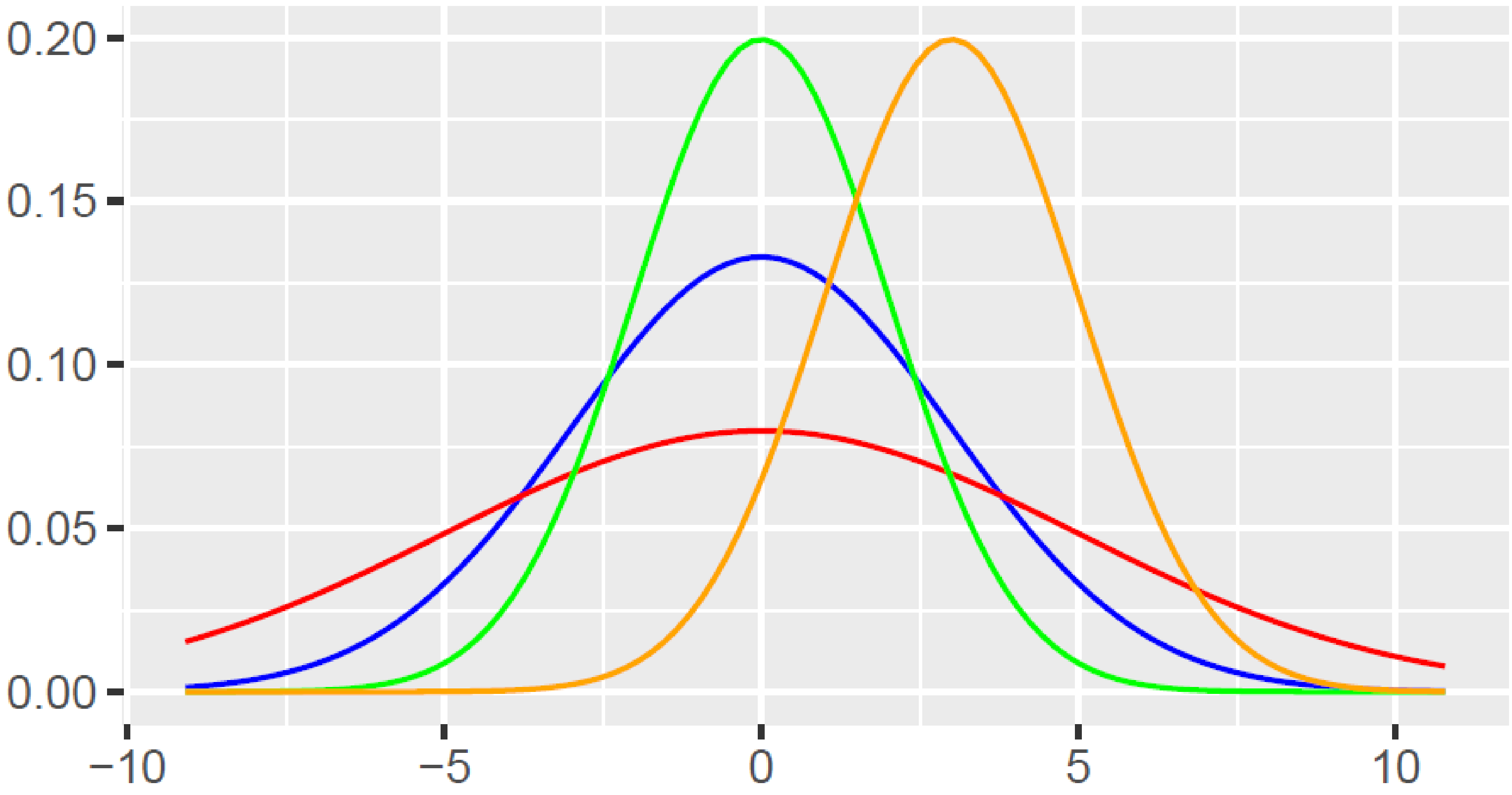
… assumptions

- **Identically distributed** - randomly drawn from the same distribution
- **Independently** - the result of one draw does not affect the others
- The distribution is assumed to be a **normal** (Gaussian) distribution of zero mean
    - Note: precisely the same distribution for every point

Normal Distribution

- Continuous bell-shaped, symmetrical distribution
- It frequently appears in nature and in industrial applications
- It is also of great theoretical importance
- It is uniquely defined by its mean and standard deviation

Example Normal Distributions

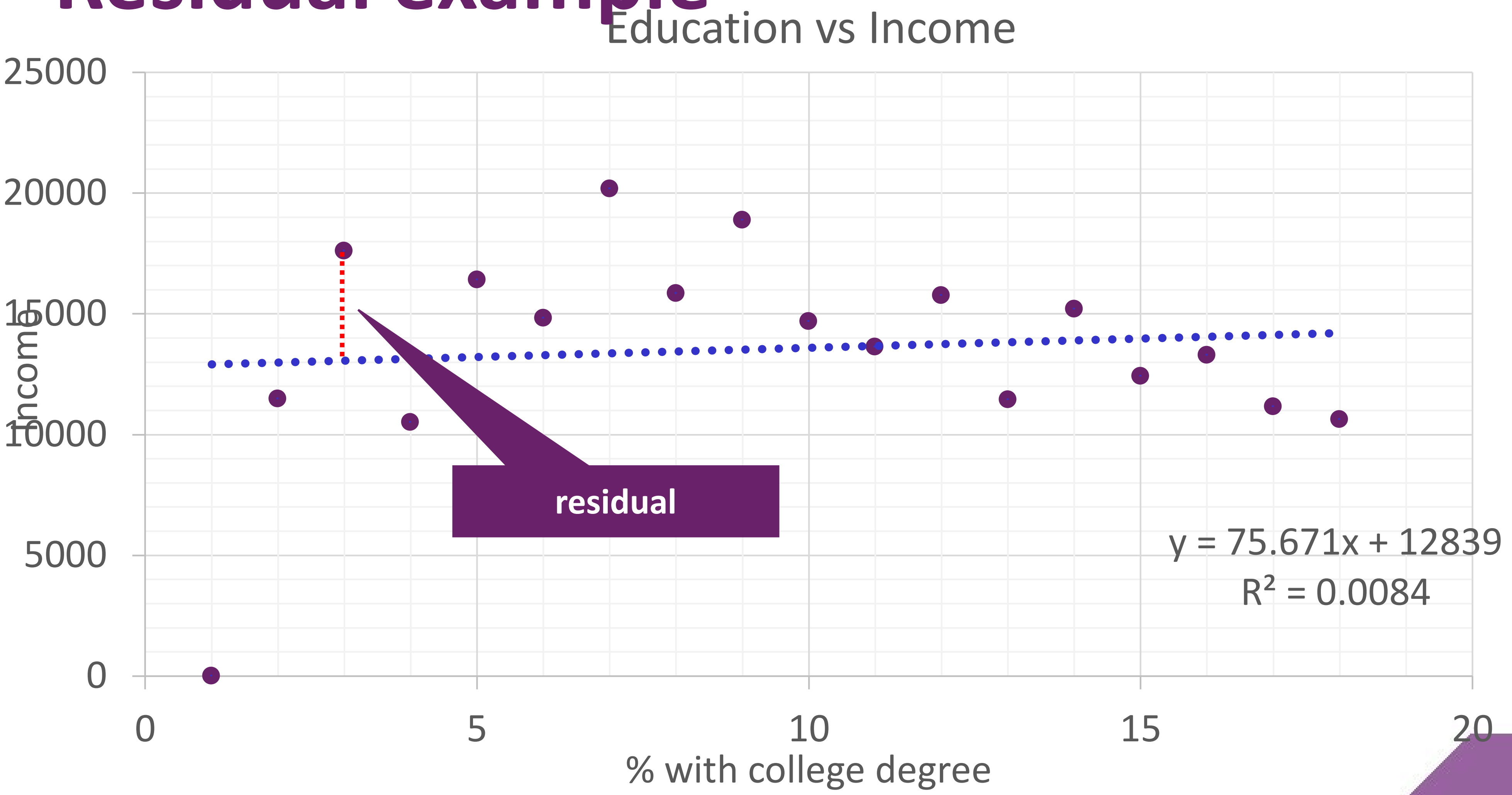


Green: $\bar{x} = 0, \sigma = 2$
Blue: $\bar{x} = 0, \sigma = 3$
Red: $\bar{x} = 0, \sigma = 5$
Orange: $\bar{x} = 3, \sigma = 2$

Residuals and Least Squares

- The *residuals* are the differences between the data values and the predictions of the linear model, i.e.
  - For a linear regression $y = \beta_0 + \beta_1 x$
  - For a data point (x_i, y_i)
    - $residual_i = y_i - (\beta_0 + \beta_1 xi)$
- The best fitting straight line is the one that minimises the sum of the squares of the residuals
  - The 'least squares' line

Residual example



How Good is Our Model?

- Coefficient of Determination R^2
  - Proportion of spread (variance) in Y explained by linear model
  - Square of Correlation Coefficient
  - Values close to 1 indicate strong linear relationship
  - Values close to 0 indicate no linear relationship

Statistical Inference

- Linear regression model follows
  - $y = \beta_0 + \beta_1 x + \varepsilon$
- Consider our data as a sample of a much larger population
- Can we reject (with x% confidence) the null hypothesis that $\beta_1 = 0$?
  - F Test (see lab)
- Find an interval estimate for the model coefficients
  - With x% confidence, each coefficient lies in a certain interval (formulae for this known).

Linear Regression Formulae

$$\boldsymbol{y = \beta_0 + \beta_1 x + \varepsilon}$$

$$\beta_1 = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

$$R^2 = \frac{\left(n\sum xy - (\sum x)(\sum y)\right)^2}{\left(n\sum x^2 - (\sum x)^2\right)\left(n\sum y^2 - (\sum y)^2\right)}$$

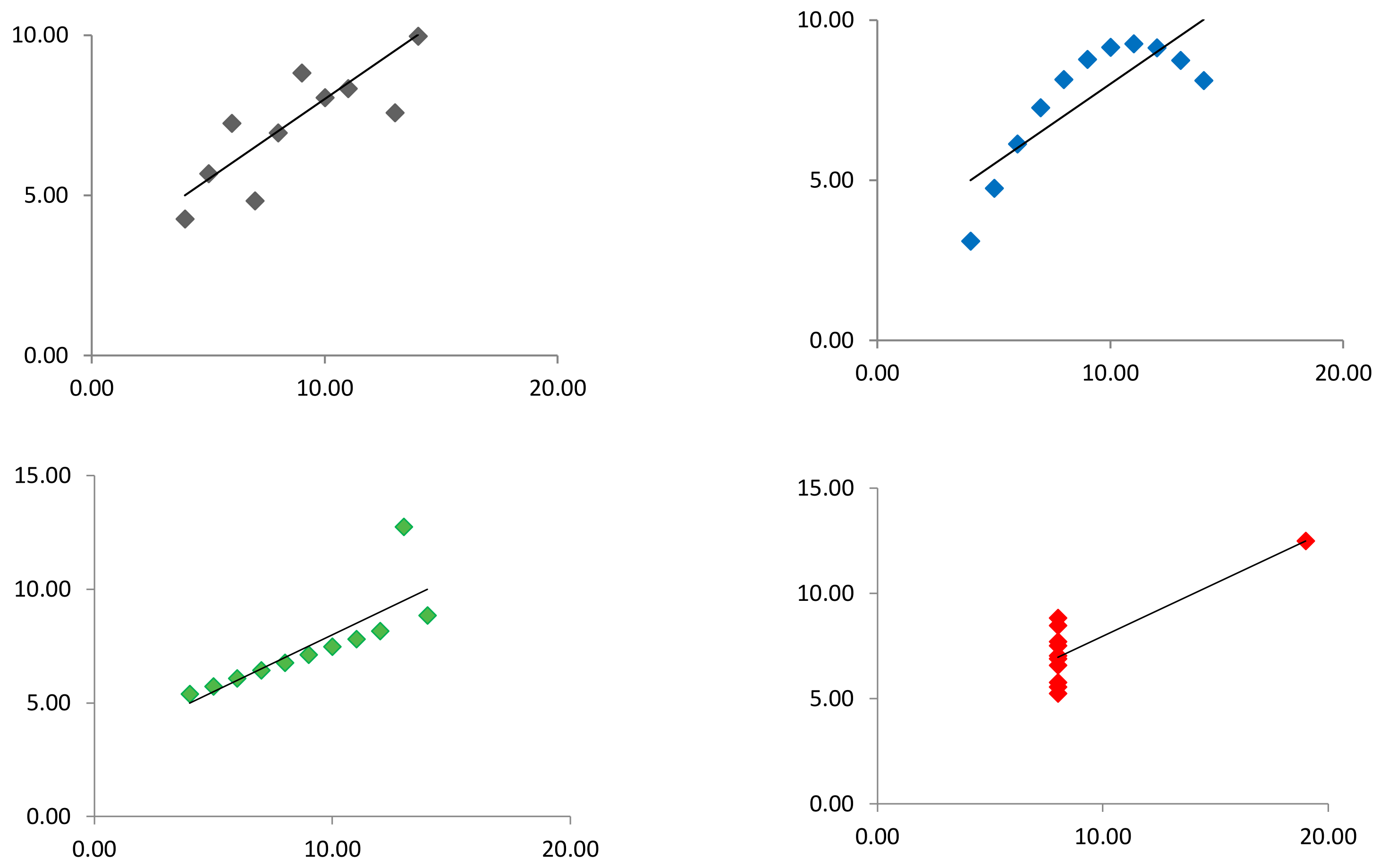
Example Calculation

- Note formulae depend on number of (pairs of) observations n and the sums
  - $\sum x, \sum y, \sum x^2, \sum y^2, \sum xy$
- Consider data for which:
  - $n = 11, \sum x = 99, \sum y = 82.5$,
  - $\sum x^2 = 1001, \ \sum y^2 = 660, \sum xy = 796.6$
  - Then $\beta_0 = 3, \ \beta_1 = 0.5, \ R^2 = 0.67$

Value is positive, but not really close to 1 so model is not ideal

- Best linear fit with moderate linear correlation
  - $\boldsymbol{y = \beta_0 + \beta_1 x + \varepsilon = 3 + 0.5\,x + \varepsilon}$
  - But, must visualise to check

Why Visualise? Visualisation shows differences!



Anscombe quartet

Anscombe’s Quartet

- Black points are what we expected
- Blue points show clear non-linear relationship
  - Should fit higher order polynomial
- Green points: outlier in Y spoils perfect fit and drags line upwards
  - Should seek to explain outlier
- Red points: outlier in X has undue influence
  - Should we be fitting this at all?

Graphing Residuals

- Since the residuals correspond to the random variable $\boldsymbol{\varepsilon}$, they should look like iind samples
- We can check this by graphing
  - Residuals vs X
    - should look random
  - Residuals vs Y
    - should look random
  - Residuals vs theoretical normal distribution.
    - should be close to straight line

Reality Checks

- Consider plausibility of linear model at outset
  - If appropriate, before collecting data
- Visualise your data
- Examine Residuals

Prediction using Linear Regression

- To predict the y-value corresponding to a particular x, we just use the straight line
- Should be interpolation – x value in range of X data used to find line
- **Extrapolation (x value outside range) is inadvisable**
- **Do not use y on x line to predict x from y!**

Generalisations

- Can fit other functions, e.g.
  - Polynomial (quadratic, cubic etc.)
  - Exponential
  - Logistic (used for probabilities)
- Multivariate Linear Regression
  - Consider multiple predictor variables

Summary

- When each data point involves 2 variables the analysis can be extended
  - Covariance $\sigma(x, y)$
  - Correlation coefficient $R(x, y)$
  - Linear regression $y = \beta_0 + \beta_1 x$
  - Coefficient of determination , R^2
- What if one of our variables is time?
  - Special case requiring separate treatment
  - Data of this type is called a **time series**
  - A series of values of the same variable at different times