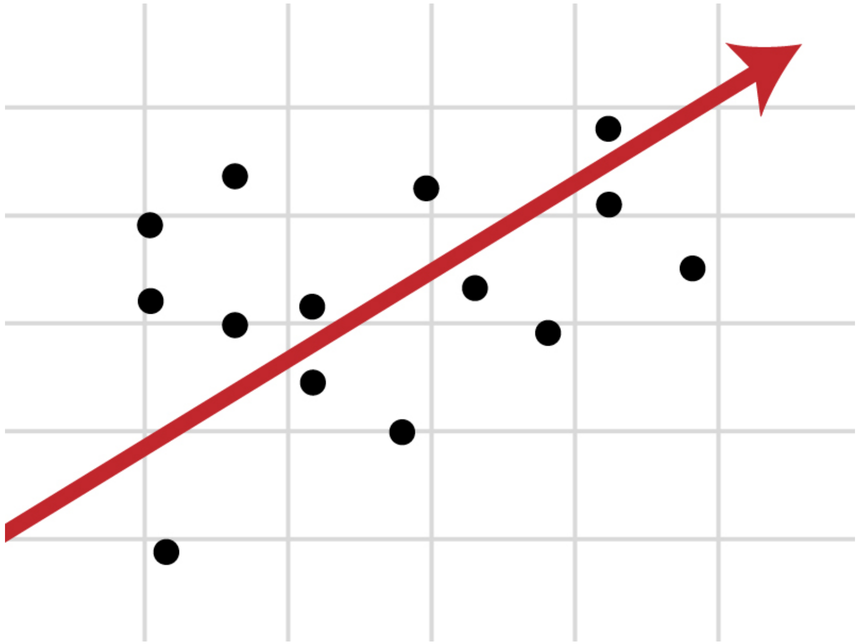
People Analytics: Strategy and Practice



Introduction to regression analysis

Modelling

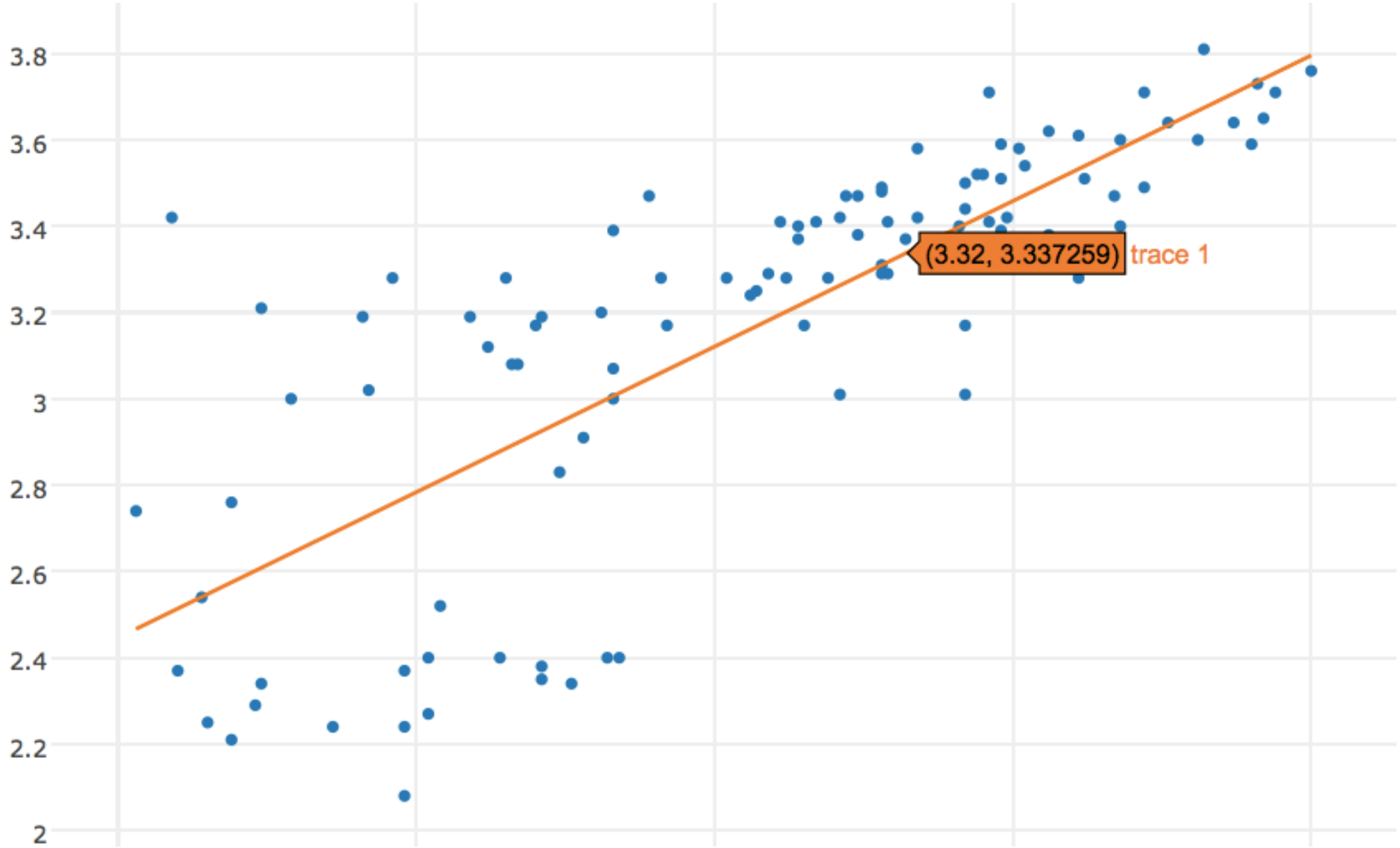
Model

Reality

Regression equation

$$Y = a + \beta_1 x_1 + \cdots \beta_n x_n + \varepsilon$$

Where Y – predicted values (dependent variable);
a - intercept; β - Beta coefficient; x - independent variable; ε – error term



Regression fit and output

- Standard error is a key for our understanding of the accuracy of predictions
- It shows how widely the data points are scattered around the regression line

$$SE_{est} = SDy\sqrt{\frac{N}{N-2}(1 - r^2)}$$

- R^2 (R squared) or the coefficient of determination serves to identify how well the regression line fits the data [0;1]

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} = 1 - \frac{\sum(y - fi)^2}{\sum(\mathrm{y} - \hat{\mathrm{y}})^2}$$

Example (I)

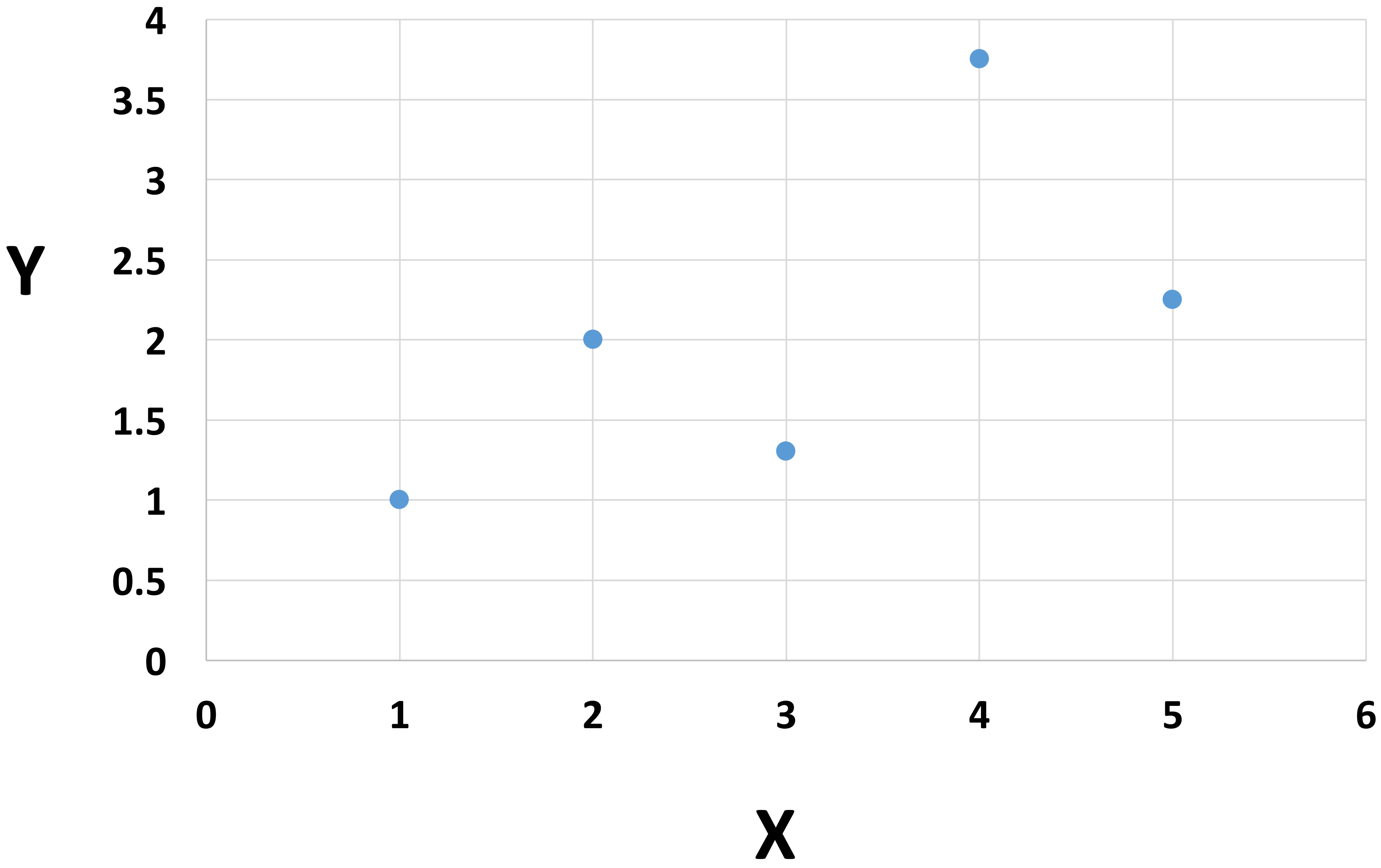
- Assume we have two random variables: X and Y

X	Y
1	1
2	2
3	1.3
4	3.75
5	2.25

- Are these two variables related to one another?

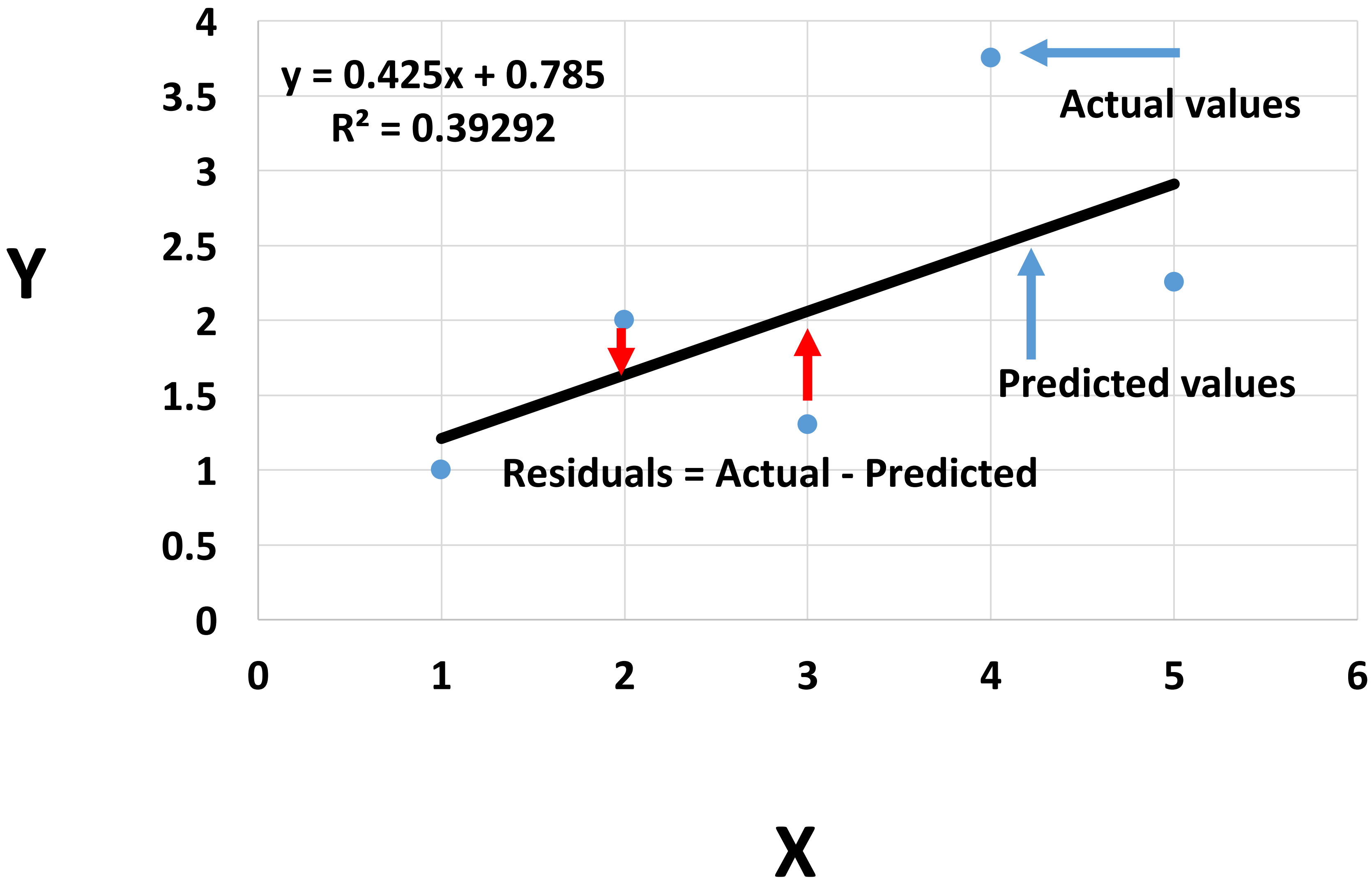
Example (II)

• Scatter plot



Example (III)

- Scatter plot



Ordinary Least Squares

- The best-fitting line in most of the cases is defined on the premise of the minimisation of the sum of the squared errors of prediction
- This procedure is termed Ordinary Least Squares (OLS)

- R^2 (R squared) or the coefficient of determination serves to identify how well the regression line fits the data [0;1]

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} = 1 - \frac{\sum(y - fi)^2}{\sum(\mathrm{y} - \hat{\mathrm{y}})^2}$$

Example (IV)

Regression parameters

X	Y	Y'	Y-Y'	$(Y\text{-}Y')^2$
1	1	1.21	-0.21	0.044
2	2	1.635	0.365	0.133
3	1.3	2.06	-0.76	0.578
4	3.75	2.485	1.265	1.6
5	2.25	2.91	-0.66	0.436

Logistic regression

Logistic regression (I)

- Assume we have a binary response outcome variable (Yes-No kind of answer)

$$P(y = 1)$$

- For example, in the WERS survey we used last time around there are 8136 union members as opposed to 13721 non-members

$$\mu = \frac{8136}{21857} = 0.372; P(y = 1) = 37.2\%$$

- What are the odds? $Odds = \frac{\pi_i}{1-\pi_i} = \frac{0.372}{1-0.372} = 0.592$

Logistic regression (II)

- Standard linear regression fails to deal with binary data
- Non-normal residuals, non-linear relationship and probabilities are discrete and locked between 0 and 1
- So we transform our model into a generalised linear one

$$Logit(\pi_i) = ln(\frac{\pi_i}{1-\pi_i}) = \beta_0 + \beta_1 X_i$$

$$ln\left(\frac{\pi_i}{1-\pi_i}\right) = [-\infty, \infty]$$

- In order to derive predicted probabilities we need to reverse the function

Logistic regression (III)

$$\left(\frac{\pi_i}{1-\pi_i}\right) = \exp(\beta_0 + \beta_1 X_i)$$

- What is the meaning of regression coefficients?
- Raw Betas show log of odds (interpretation differs slightly for categorical and continuous predictors)
- It is easy to get odds from log odds and then to derive probabilities