Linear Regression

Resources

- Machine Learning for Hackers, Drew Conway & John Myles White, O'Reilly, Chapter 5.
- http://wweb.uta.edu/insyopma/baker/
- http://www.r-tutor.com/elementary-statistics

Models

- Deterministic Model: an equation or set of equations that allow us to fully determine the value of the dependent variable from the values of the independent variables.

  Area of a circle: $A = \pi * r^2$

- Probabilistic Model: a method used to make predictions but the predictions will be approximate due to the ***randomness*** that is part of a real-life process.
- For example, the price of a house (y) depends on the size of the house (x)
  - $y = 25{,}000 + 75x$

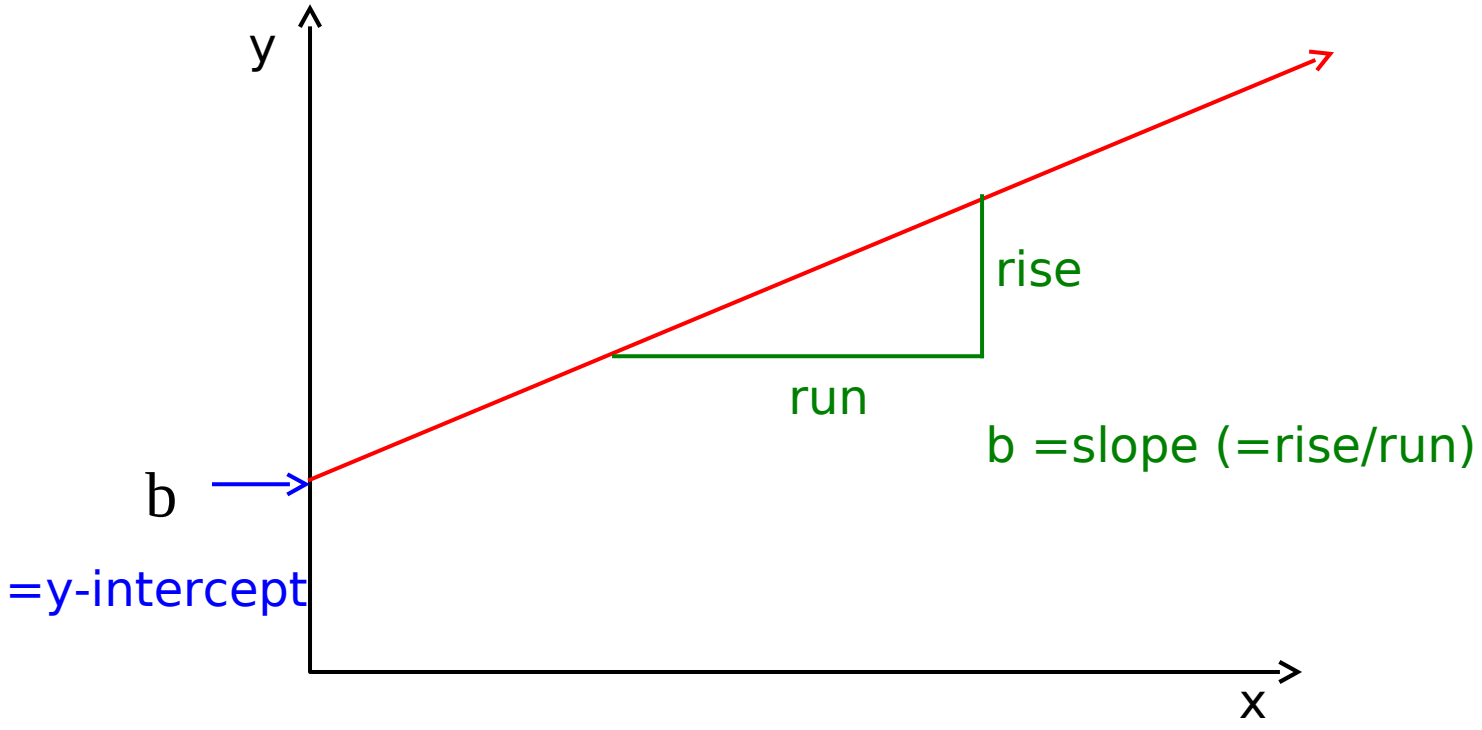
Linear Regression

- Regression analysis is used to predict the value of one variable (the ***dependent variable***) on the basis of other variables (the ***independent variables***).
- Dependent variable: denoted y
- Independent variables: denoted x_1, x_2, …, x_k
- Simple Linear Regression – one independent variable.

Simple Linear Regression

- Meaning of a and b
  - a > 0 [positive slope]
  - a < 0 [negative slope]

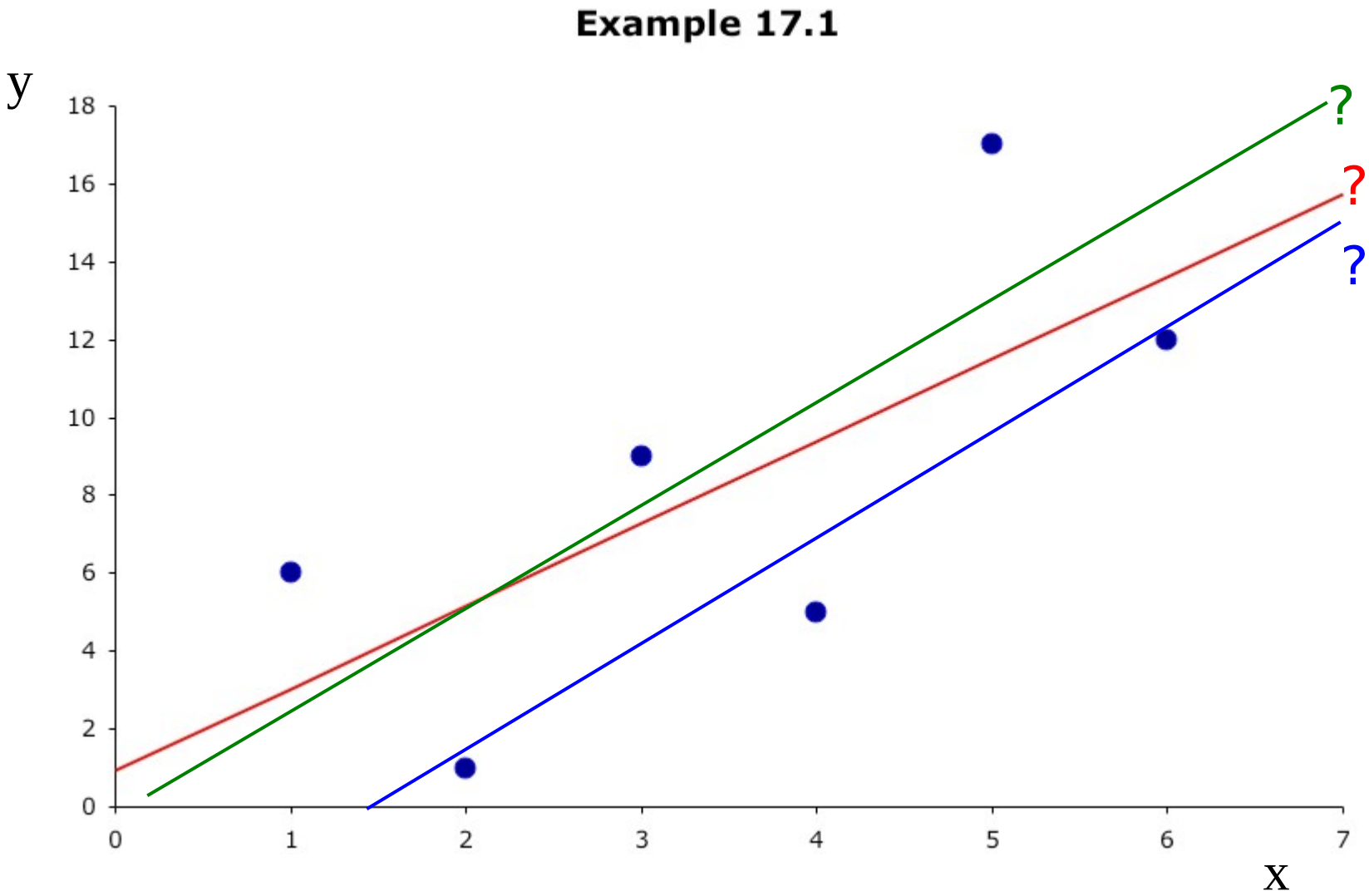
$$y = a\,x + b$$



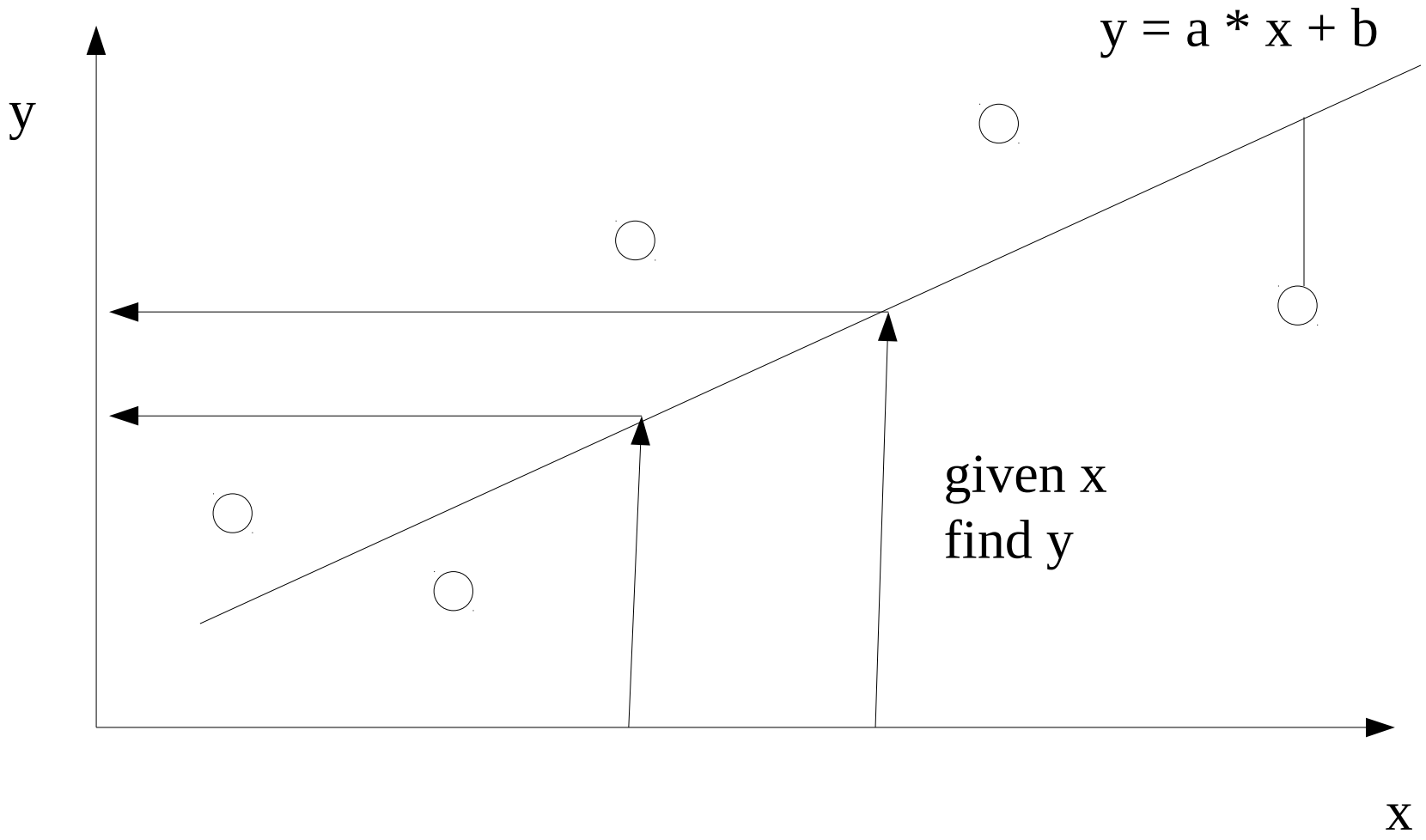
Find the Parameters a and b

- Normally we have a set of existing data, i.e. corresponding values of x and y.
- From there we want to estimate a and b.
- This corresponds to finding the line of best fit through the points.

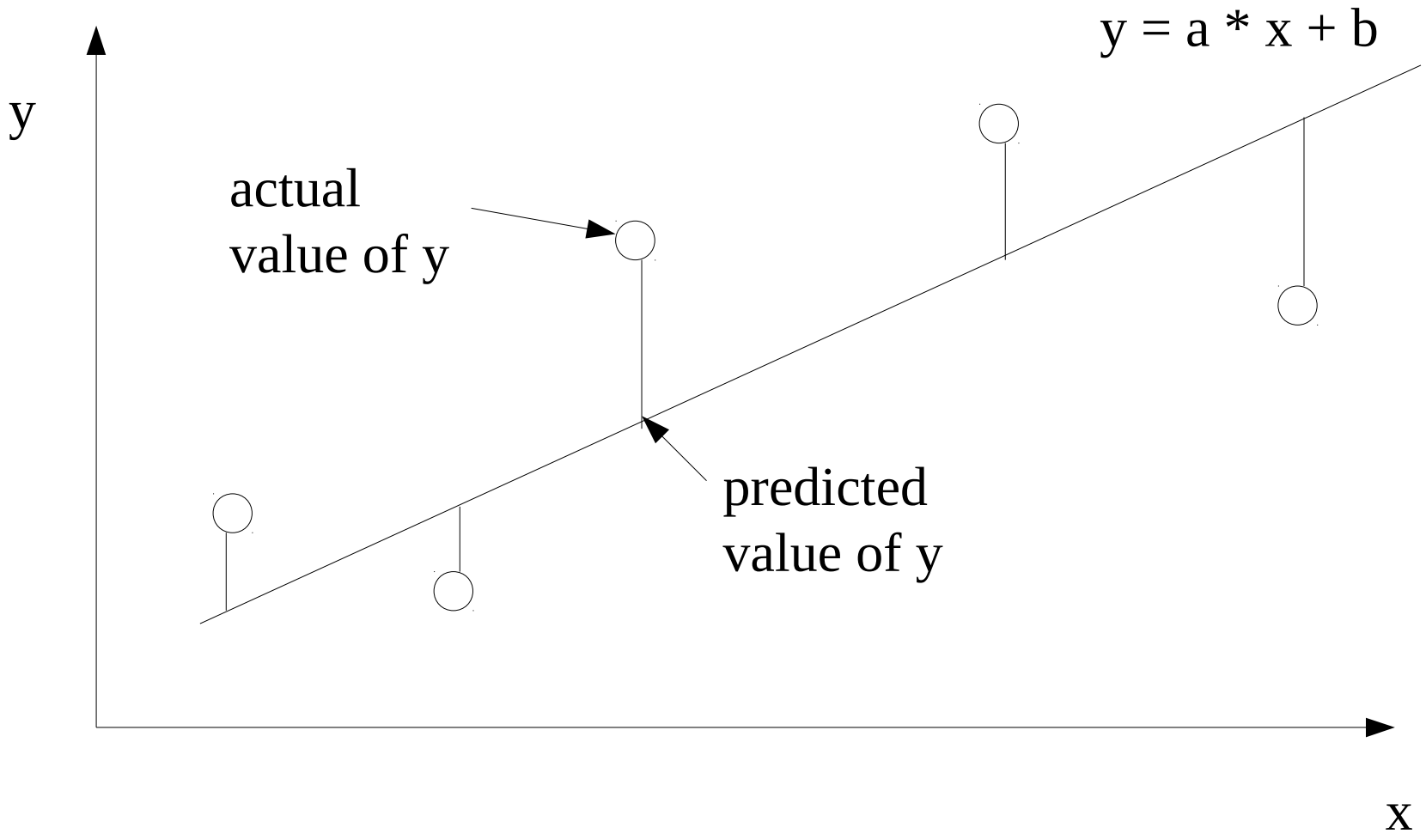
Which line has the best “fit” to the data?



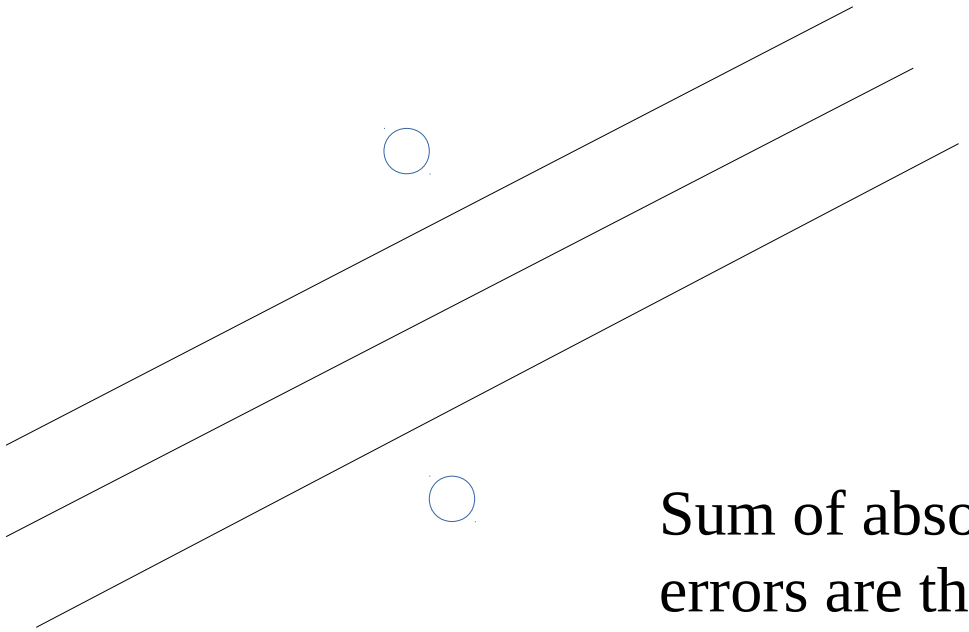
Prediction



Errors



Errors (Residuals)



Sum of absolute value of errors are the same for all three lines

Finding Line of Best Fit

- Minimize sum of the errors?
  - No, positive and negative errors will cancel
- Minimize sum of absolute value of the errors?
  - No, when we move a line up and down this sum can stay the same.
- Minimize the sum of the squares of the errors.
  - Yes.
- [Errors, often known as residuals.]

Estimating the Coefficients Using the Least Squares method

- The least squares method produces a straight line that minimizes the sum of the squares of the errors.
- Errors (or Residuals)
  - differences between the points and the line
  - differences between predicted value of y and actual value of y.
  - Predicted value of y often written as $\hat{y}$

Objective Function

- This is an opimization problem.
- An opimization problem is normally solved by determining values that minimizes (or maximizes) an objective function.
- For simple linear regression, the objective function is the sum of the squares of the errors (residuals)
- Also known as a cost function.