Equation of a line

WiSe23/24

Daniela Palleschi

2023-10-10

Table of contents

- Learning Objectives . . . 1
- Resources . . . 2
  - Statistical tests versus models . . . 2
- (Linear) Regression . . . 2
  - Types of regression . . . 2
- Straight lines . . . 2
  - A line = intercept and slope . . . 3
  - Equation of a line . . . 3
  - Intercept (b_0) . . . 4
  - Slopes (b_1) . . . 4
- Error and residuals . . . 5
  - Method of least squares . . . 8
- Important terms . . . 8
- Exercise . . . 9
  - Pen-and-paper . . . 9

Learning Objectives

Today we will learn…

- the equation of a line
- about intercepts, slopes, and residuals

regression type	predictor	outcome
simple regression	Single predictor	continuous (numerical)
multiple regression	multiple predictor	continuous (numerical)
hierarchical/linear mixed models/linear mixed effect models	include random effect	continuous (numerical)
generalised linear (mixed) models: logistic regression	as above	binary/binomial data
generalised linear (mixed) models: poisson regression	as above	count data

Resources

- relevant readings:
  - Winter (2013)
  - Winter (2019) (Ch. 3)

Statistical tests versus models

- you're probably familiar with statistical tests like the *t*-test or Chi-squared test
- however, common statistical tests are simplified linear models (see also Statistical tests vs. linear regression)
  - but without the added power of linear models (e.g., multiple predictors, crossed random effects)
- statistical tests tell us something about our data
- statistical *models* can generalise beyond our data

(Linear) Regression

- we need to fit a model to our data to make predictions about hypothetical observations
  - i.e., to *predict* values of our outcome/response variable based on one (or more) predictor variables
- this model can then *predict* values of our DV based on one (or more) IV(s), i.e., *predicting* an outcome variable - because we're making predictions, we need to take into account the variability (i.e., *error*) in our data
- but how do we fit these models, and what does that even mean?

Types of regression

Straight lines

- *linear regression* summarises the data with a straight line

    - we *model* our data as/fit our data to a straight line

- *straight lines* can be defined by
    - Intercept (b_0)
        * value of Y when $X = 0$
    - Slope (b_1)
        * gradient (slope) of the regression line
        * direction/strength of relationship between x and y
        * regression coefficient for the predictor
- so we need to define an intercept and a slope

A line = intercept and slope

- a line is defined by its intercept and slope
    - in a regression model, these two are called `coefficients`

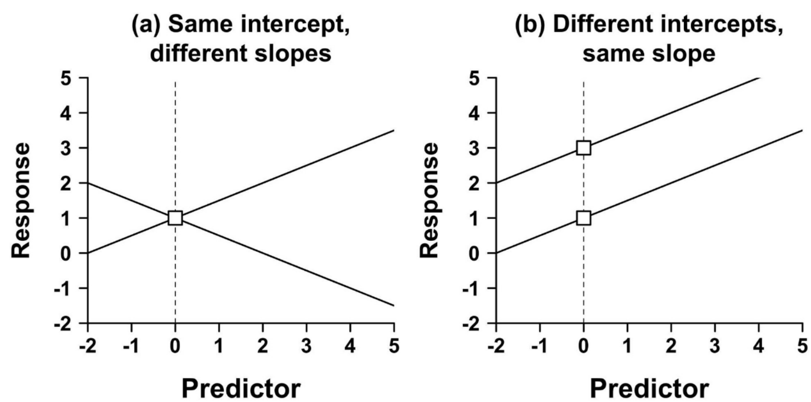


Figure 4.2. (a) Two lines with positive and negative slopes that go through the same intercept; (b) two lines with the same positive slope that have different intercepts

Figure 1: Image source: Winter (2019) (all rights reserved)

Equation of a line

- the following are all different ways to say that a value of y for a given value of x (indicated by $_i$) is equal to the *intercept* (b_0) plus the *slope* (b_1) multiplied by the value of x

$$

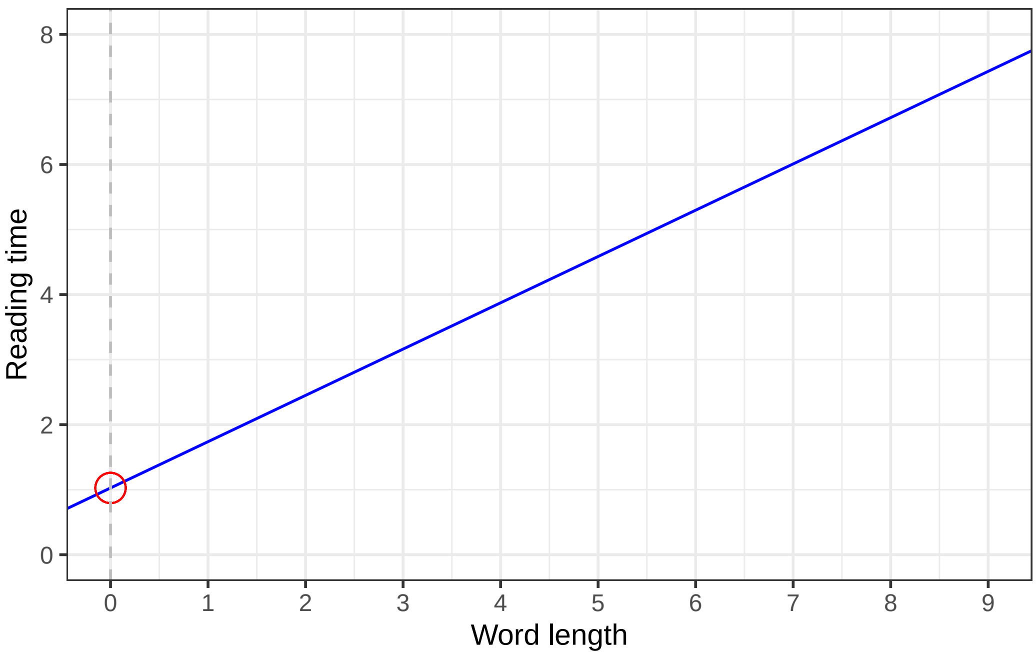
y & = mx + c\ Y_i & = b_0 + b_1X_i \ outcome_i & = (model) \ y_i & = intercept + slope*x_i

$$

- with this equation, we can *predict* values of y (our outcome variable) for a given value of x (our predictor variable)

Intercept (b_0)

- the value of y when $x = 0$



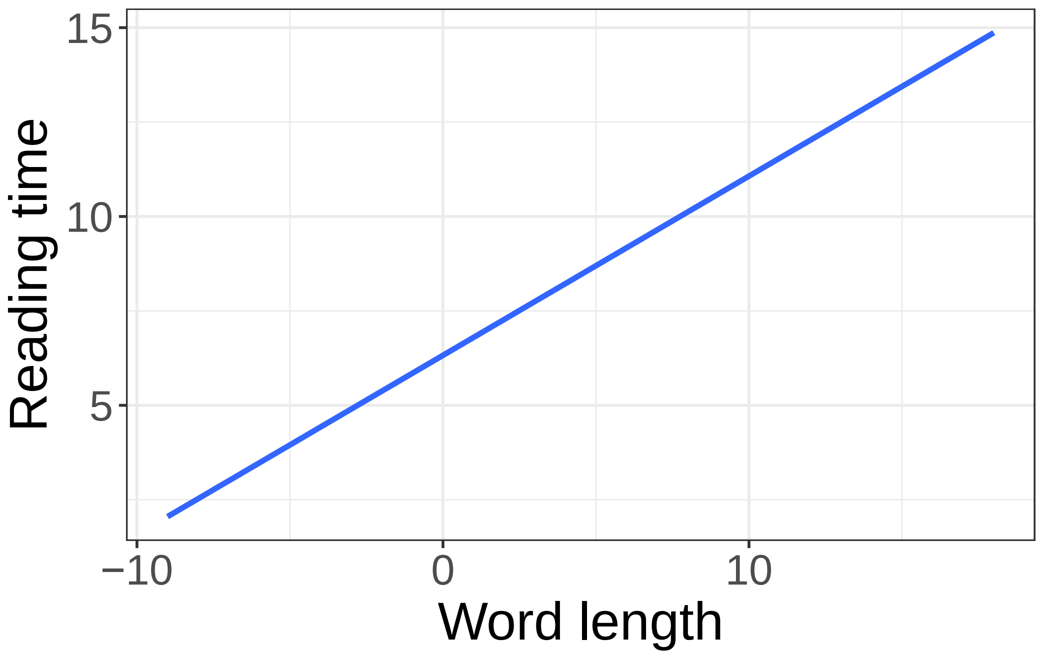
Slopes (b_1)

- slopes describe a change in y (Δy) over a change in x (Δx)
  - positive slope: as x increases, y increases
  - negative slope: as x increases, y decreases
  - if the slope is 0, there is no change in y as a function of x
- or: the change in y when x increase by 1 unit

  - sometimes referred to as "rise over run": how do you 'rise' in y for a given 'run' in x?

$$slope = \frac{\Delta y}{\Delta x}$$

- what is the intercept of this line?
- what is the slope of this line?

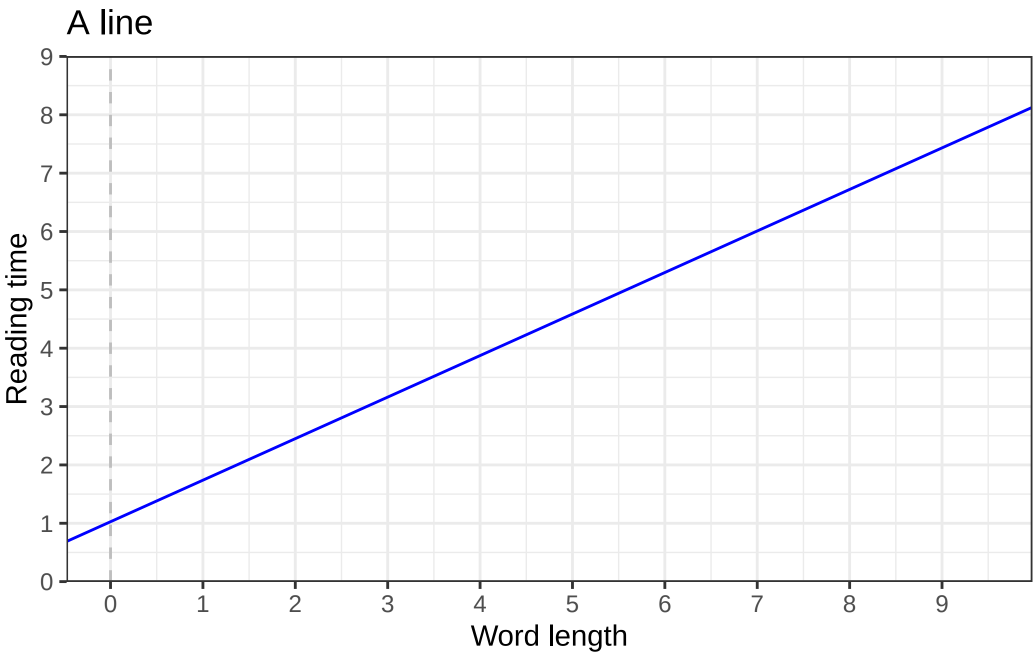


Error and residuals

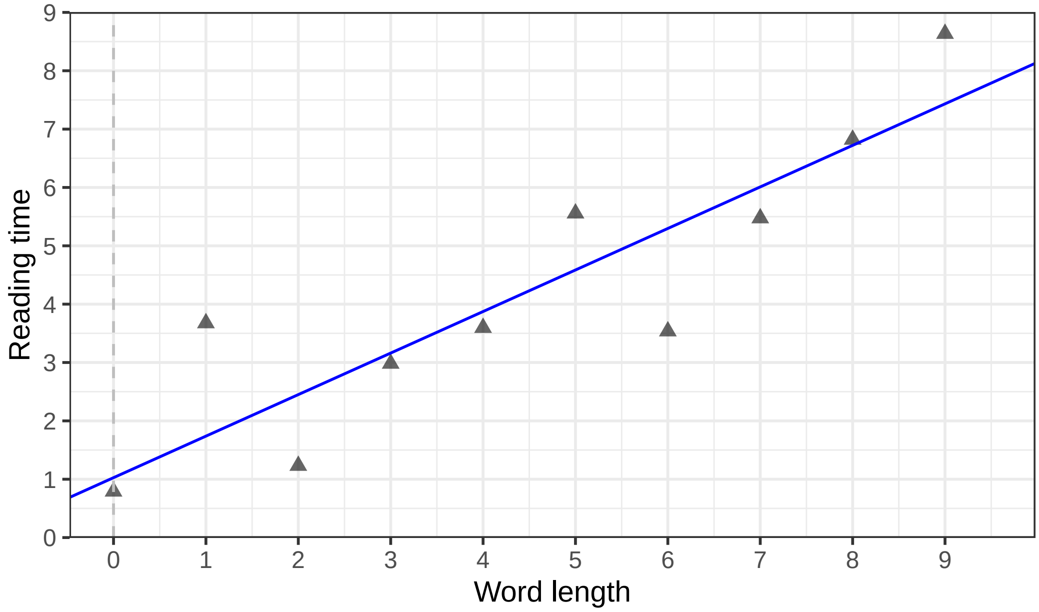
- *fixed effects* (IV/predictors): things we can understand/measure
- *error* (random effects): things we cannot understand/measure
  - in biology, social sciences (and linguistic research), there will always sources of random error that we cannot account for
  - random error is less an issue in e.g., physics (e.g., measuring gravitational pull)
- *residuals*: the difference (vertical difference) between **observed data** and the **fitted values** (predicted values)

Equation of a line

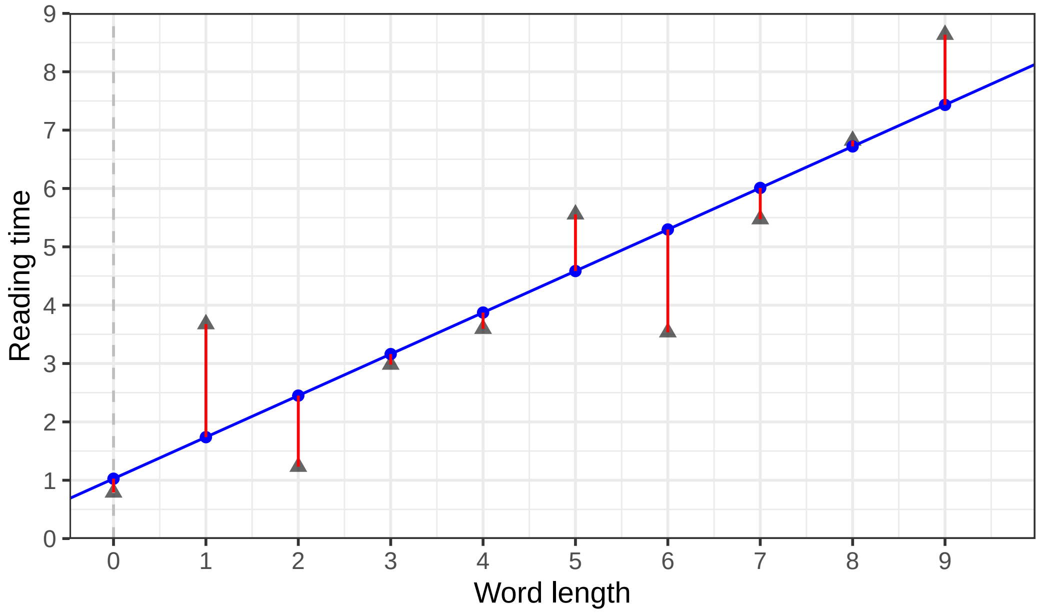
$$
y & = mx + c\ Y_i &= (b_0 + b_1X_i) + _i\ outcome_i & = (model) + error_i\
y_i & = (intercept + slope*x_i) + error_i
$$



A line with data points



A line with data points and regression line



Method of least squares

- so how is any given line chosen to fit any given data?
- the ***method of least squares***
  - take a given line, and square all the residuals (i.e., $residual^2$)
  - the line with the lowest ***sum of squares*** is the line with the best fit to the given data
  - why do we square the residuals before summing them up?
    - so all values are positive (i.e., so that negative values don't cancel out positive values)
- this is how we find the ***line of best fit***
  - R fits many lines to find the one with the best fit

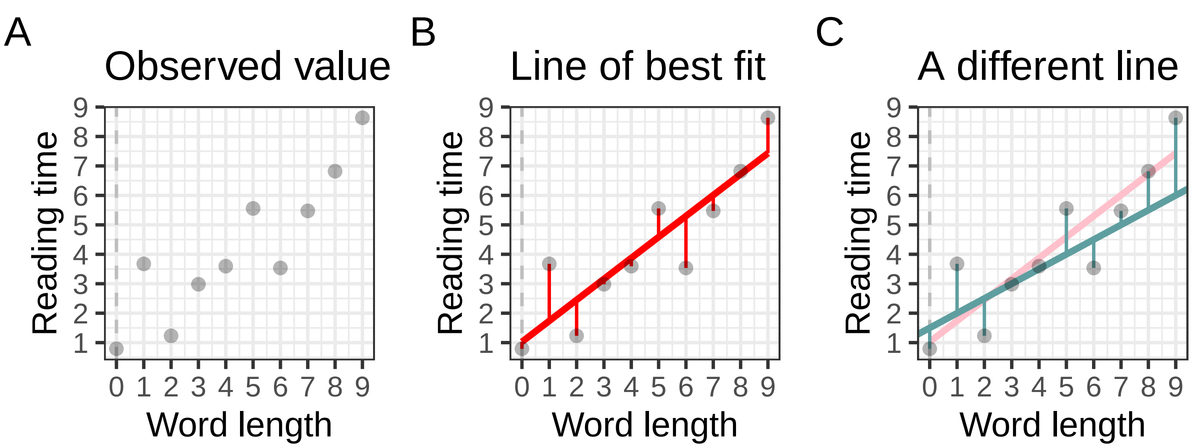


Figure 2: Observed values (A), Residuals for line of best fit (B), A line of worse fit with larger residuals (C)

Learning Objectives

Today we learned…

Important terms

Term	Definition	Equation/Code
Intercept	Value of y for x=0	b0

Exercise

Pen-and-paper

You will receive a piece of paper with several grids on it. Follow the instructions, which include drawing some lines.

Literaturverzeichnis

Winter, B. (2013). *Linear models and linear mixed effects models in R: Tutorial 1.*

Winter, B. (2019). Statistics for Linguists: An Introduction Using R. In *Statistics for Linguists: An Introduction Using R.* Routledge. https://doi.org/10.4324/9781315165547