CSCI 3022

intro to data science with probability & statistics

November 9, 2018

Introduction to statistical regression

- git pull
- arkaive

Department of Computer Science
UNIVERSITY OF COLORADO BOULDER

Dan Larremore

Stuff & Things

- **HW5** due today. Giddyup!

Today: linear regression.

- **Examples**:
  - given a person's age and gender, predict their height.
  - given the square footage and number of bathrooms in a house, predict its sale price.
  - given unemployment, inflation, number of wars, and economic growth, predict the president's approval rating.
  - given a user's browsing history, predict how long they will stay on a product page.
  - given the advertising budget expenditures in various media markets, predict the number of products sold.

Today, we start in the notebook

- Pull that in-class notebook, and let's get started!

Simple Linear Regression Model

- **Definitions and Assumptions** of the simple [one independent variable] linear regression model:

1. $y_i = \underbrace{\alpha + \beta x_i}_{\text{linear}} + \underbrace{\varepsilon_i}_{\text{noise}}$ true underlying rel'ship is $y = \alpha + \beta x$
2. Each ε_i is drawn indep. from the same distr. I.I.D.
3. $\varepsilon_i \sim N(0, \sigma^2)$

   Key: mean is zero.

SLR Model

- **Vocabulary** for the SLR model:
- ***X***: the **independent variable**, the **predictor**, the **explanatory variable**, the **feature**.
  - *X is not random!*
- ***Y***: the **dependent variable**, the **response variable**.
  - For a fixed *x*, Y is *random.*
- **ε**: the **random deviation** or **random error** term.
  - For a fixed *x*, ε *is random*.

What exactly is ε doing?

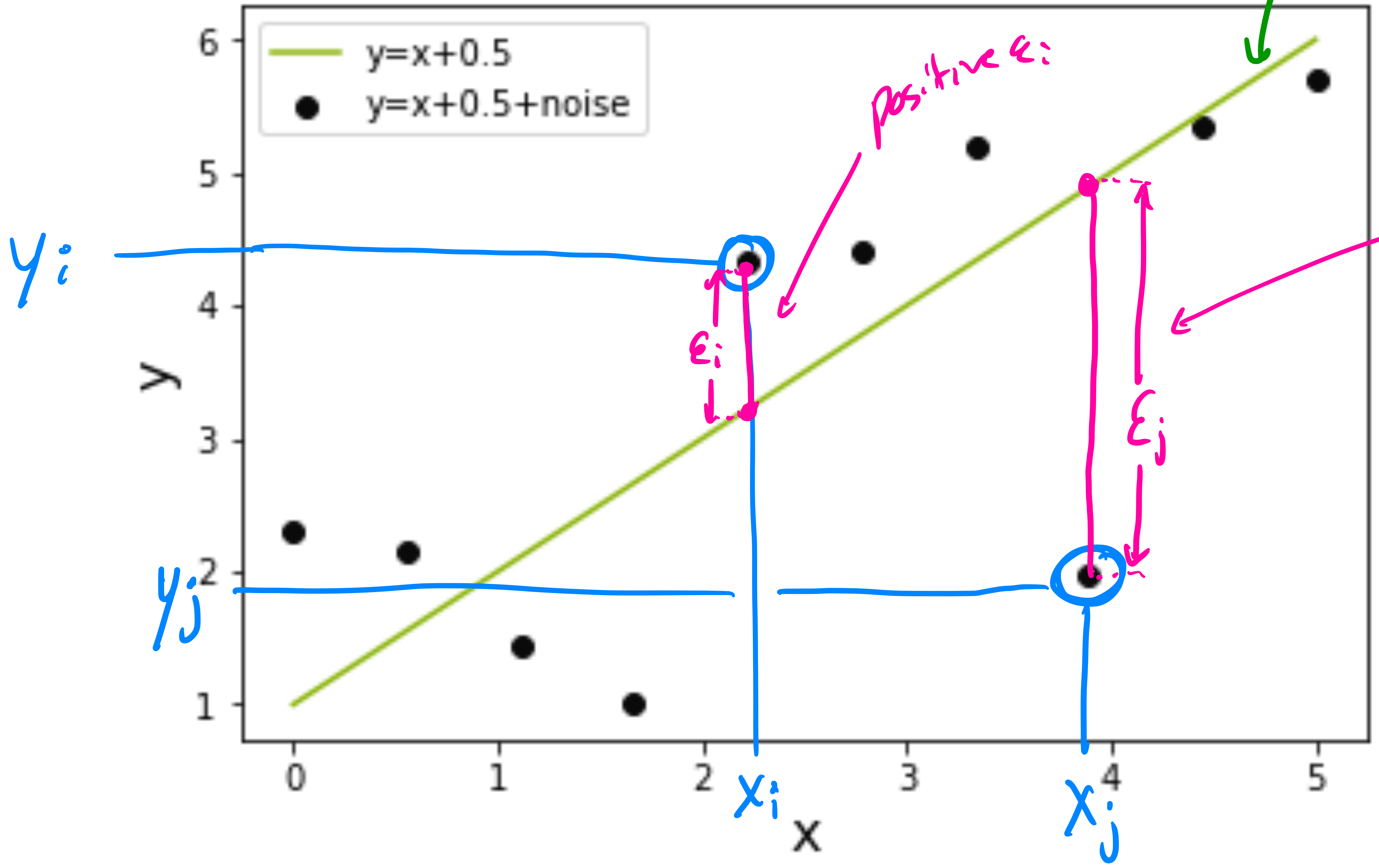
$$y_i = \overbrace{\alpha + \beta x_i}^{\text{fixed}} + \varepsilon_i$$

random

SLR Model

true regression line
$y = \alpha + \beta x$

- The points $(x_1,y_1),\ldots,(x_n,y_n)$ resulting from *n* independent observations will then be scattered about the true regression line:

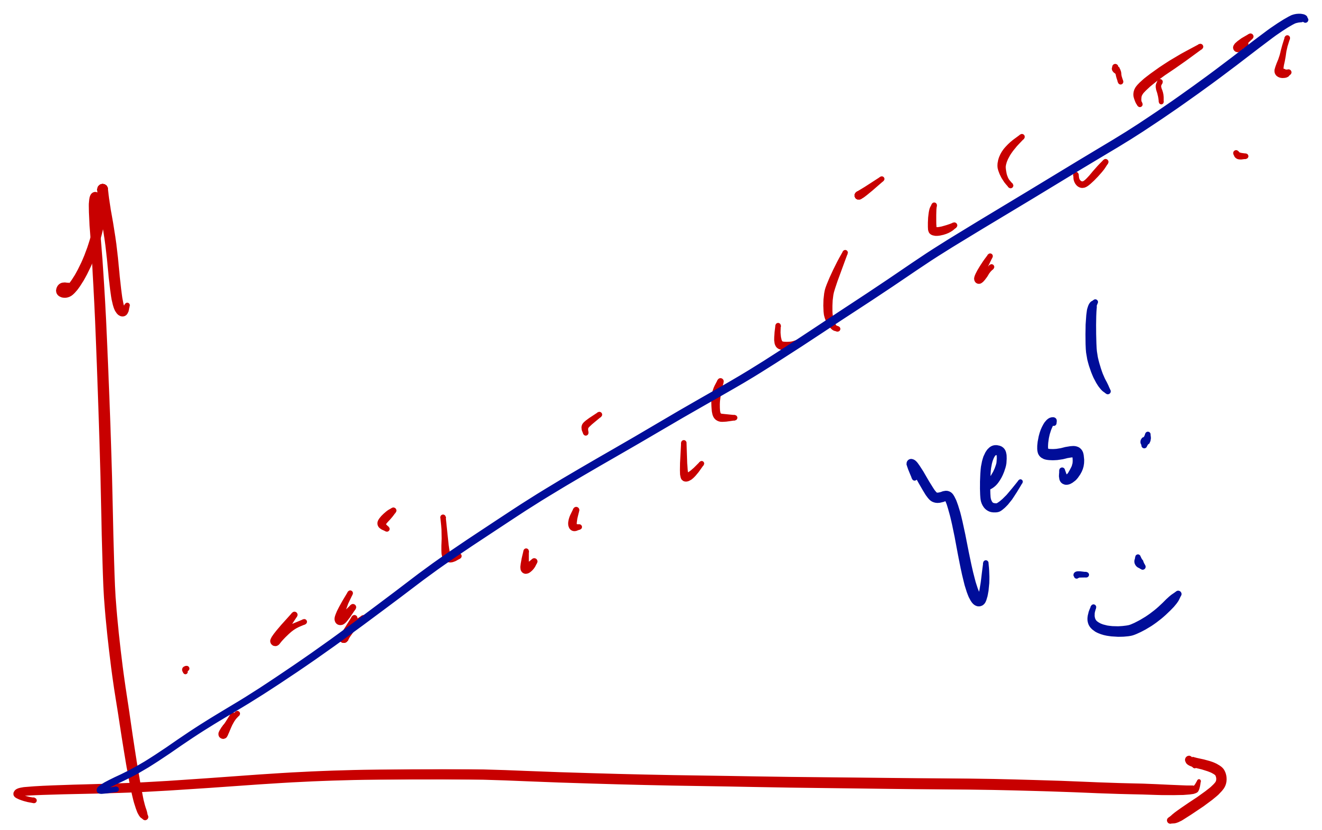


SLR: theory

- How do we know that a simple linear regression is appropriate?
- Theoretical considerations

  ① Belief or knowledge about where the noise comes from in my real application.

  ② Relationship btwn X and Y.
- Scatterplots
- Knowledge of the process generating the data.



SLR Model

$Y = \alpha + \beta x + \varepsilon$

$E[Y] = E[\alpha + \beta x + \varepsilon]$

$= \alpha + \beta x + E[\varepsilon]$

$= \alpha + \beta x$

recall

$\varepsilon \sim N(0, \sigma^2)$

$\uparrow$

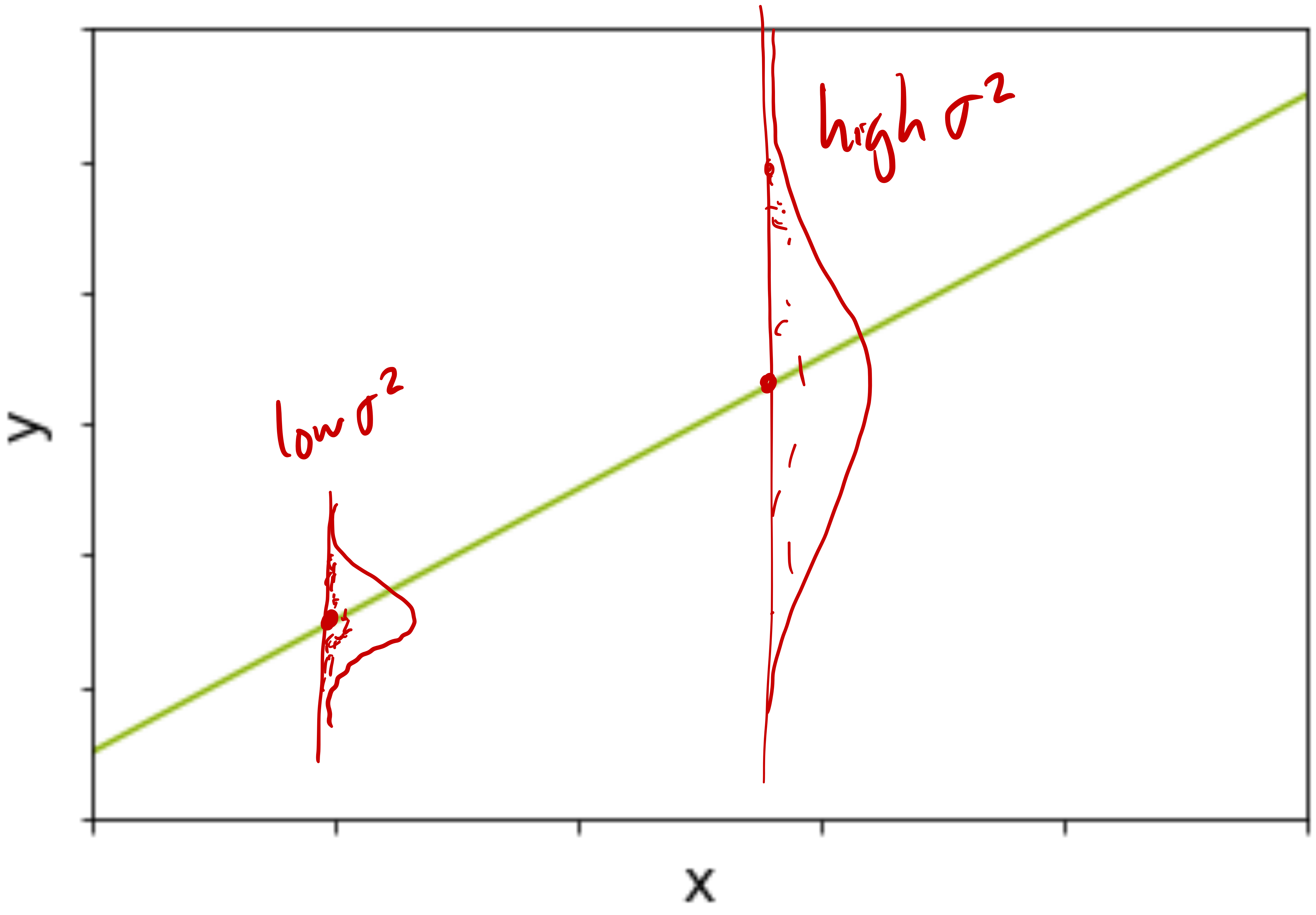
$E[\varepsilon]$

- **Interpreting parameters**:

- Y is a random variable. What is its expectation, E[Y]? $E[Y] = \alpha + \beta x$
- α (the intercept of the true regression line):
  - The average value of Y when x is zero. This is sometimes called the **baseline average**.
- β (the slope of the true regression line):
  - The average change in Y associated with a 1-unit increase in the value of x.

The Error Term

$\varepsilon_i \sim N(0, \sigma^2)$



- The variance parameter σ^2 determines the extent to which each normal curve spreads out about the regression line.

Generative model vs regression

- So far, we've written down a **generative model** where we choose **parameters** and then **generate data stochastically**.
- But really, we want to run this process in reverse. We have data, and we want to **find/learn/estimate the parameters** that explain the data.

GM + parameters $\xrightarrow{\text{stochastic draw}}$ Data

GM + parameters $\xleftarrow{\text{inference}}$ Data

want

How can we estimate model parameters?

- Plan of attack: the variance of our model σ^2 will be smallest if the differences between the estimate of the true line and each point is the smallest. **This is our goal: minimize** σ^2

  recipe

- We use our sample data, which consists of *n* observed (x,y) pairs to *estimate* the regression line. $(x_1,y_1), \ldots, (x_n,y_n)$ ingredients.

  goal: cook up $\hat{\alpha}, \hat{\beta}$

- What are we assuming about each of the data pairs?

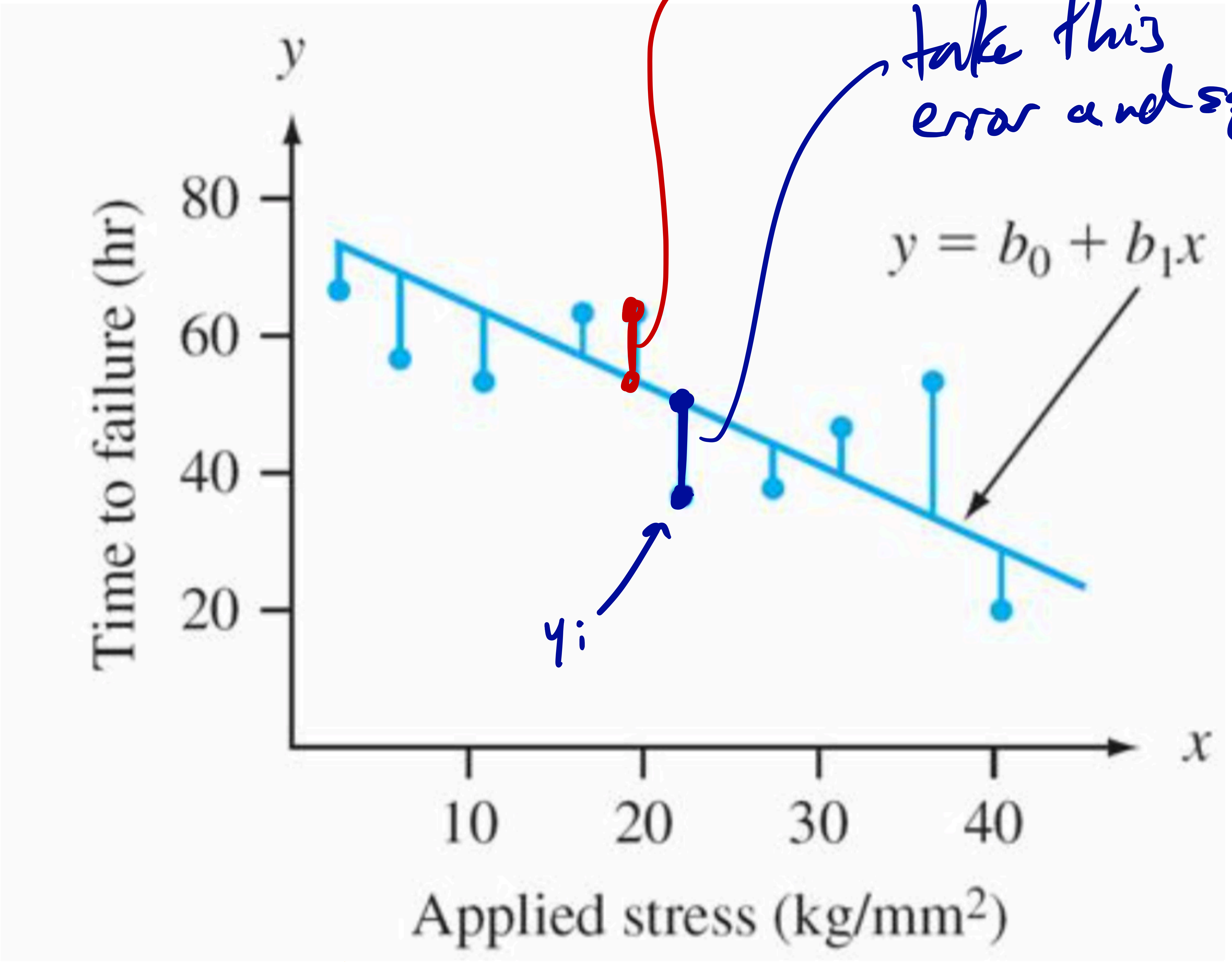
  Indep. of errors ε_i, ε_1 has no bearing on $\varepsilon_2, \varepsilon_3, \ldots$

Estimating model parameters

- The **best fit line** is motivated by the principle of least squares, which can be traced back to the German mathematician **Gauss** (1777– 1855):.
- A line provides the best fit to the data if the sum of the squared vertical distances (deviations) from the observed points to that line is as small as it can be.

square this too.

take this error and square it



square each error...

add them all up!

Estimating model parameters

- The sum of the squared deviations (also called errors) from the points (x_1,y_1), … , (x_n,y_n) to the line is then

$$\underbrace{SSE(\alpha,\beta)}_{\text{sum of squared errors}} = \sum_{i=1}^{n} \Big(\underbrace{y_i}_{\text{actual y-value}} - \underbrace{(\alpha + \beta x_i)}_{\text{y value that I predict (on the line)}} \Big)^2$$

- The “point estimates” of the slope and intercept parameters are called the **least squares estimates**, and are defined to be the values that minimize the SSE.

Find α, β to minimize $SSE(\alpha,\beta)$

Estimating model parameters

- The **fitted regression line** or **least squares line** is then the line whose equation is:

$y = \hat{\alpha} + \hat{\beta} x$

\hat{\alpha}

hat means "best fit" parameters.

- The minimizing values of α and β are found by taking [partial] derivates of SSE with respect to α and β , setting each equal to zero, and solving.
- [Take a derivative and set=0? Sounds like calculus!]

Calc III

Estimating model parameters

partial ∂

$$SSE(\alpha,\beta) = \sum_{i=1}^{n}\left(y_i - (\alpha + \beta x_i)\right)^2$$

$$\frac{\partial SSE(\alpha,\beta)}{\partial \alpha} = \sum_{i=1}^{n} \frac{\partial}{\partial \alpha}\left(y_i - (\alpha + \beta x_i)\right)^2$$

$$= \sum_{i=1}^{n} \cancel{2}\left(y_i - (\alpha + \beta x_i)\right)\cancel{(-1)} = 0$$

set to zero to minimize.

$$\Rightarrow \sum_{i=1}^{n} y_i - \alpha - \beta x_i = 0$$

$$= \sum_{i=1}^{n} y_i - \underbrace{\sum_{i=1}^{n} \alpha}_{= -n\alpha} - \beta \sum_{i=1}^{n} x_i = 0$$

$$\Rightarrow \frac{1}{n}\sum_{i=1}^{n} y_i - \beta \frac{1}{n}\sum_{i=1}^{n} x_i \quad -\frac{1}{\cancel{n}}\cancel{n}\alpha = 0$$

same procedure.

$$\boxed{\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

1

$(x_i - \bar{x})(x_i - \bar{x})$

2

$$\boxed{\bar{y} - \hat{\beta}\bar{x} = \hat{\alpha}}$$

Estimating model parameters

no.

Does it work?

- Let’s dig into problem 2 in the in-class notebook to see how this works.

Residuals

- The **fitted** or **predicted** values __________________ are obtained by substituting $x_1,\ldots x_n$ into the equation of the estimated regression line.

- The **residuals** are the differences between the observed and fitted *y* values:

Residuals

- Why are the residuals estimates of the error?

Maximum likelihood estimates

- Rather than minimizing the sum of the squared errors to find the parameters of the model, we can *maximize the likelihood of the data* by changing the parameters.
- You already know **maximum likelihood estimates** but we never called them that before.
- Imagine that we flip a biased coin and get 5 heads and 1 tails. What is the maximum likelihood estimate of the coin's bias, p?

Maximum likelihood estimates

- Three steps:
    1. Assume the parameter *p* is fixed (for now).
    2. What is the probability that we observe 5H and 1T, given p? Note: this probability is called *the likelihood*. If we take a log, this is now called the *log likelihood*.
    3. Take the derivative of step 2 with respect to p and set equal to zero. In other words, maximize the likelihood of getting 5H and 1T by finding the optimal p.

Maximum likelihood estimates

Maximum likelihood estimates

- We can repeat these steps for the linear regression problem.