CSCI 3022

intro to data science with probability & statistics

Lecture 22

April 6, 2018

Introduction to statistical regression

CI for std. dev of Normally distr. draws is indeed

$$\sqrt{\frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}}} < \sigma < \sqrt{\frac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}}}$$

slides are correct :)

Department of Computer Science
UNIVERSITY OF COLORADO BOULDER

Dan Larremore

Stuff & Things

- **HW5** due today. Giddyup!

Thx
Google Image

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

aka. features predictors

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

one x

1. $y_i = \underbrace{\alpha + \beta x_i}_{\text{linear}} + \underbrace{\varepsilon_i}_{\text{noise}}$ true underlying relationship is $y = \alpha + \beta x$

2. Each ε_i is drawn independently from same distr. IID

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

key! mean is 0

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*.

not random

$$y_i = \alpha + \beta X_i + \varepsilon_i$$

random

random

What exactly is ε doing?

SLR Model

$\alpha = 0.5$

$\beta = 1$

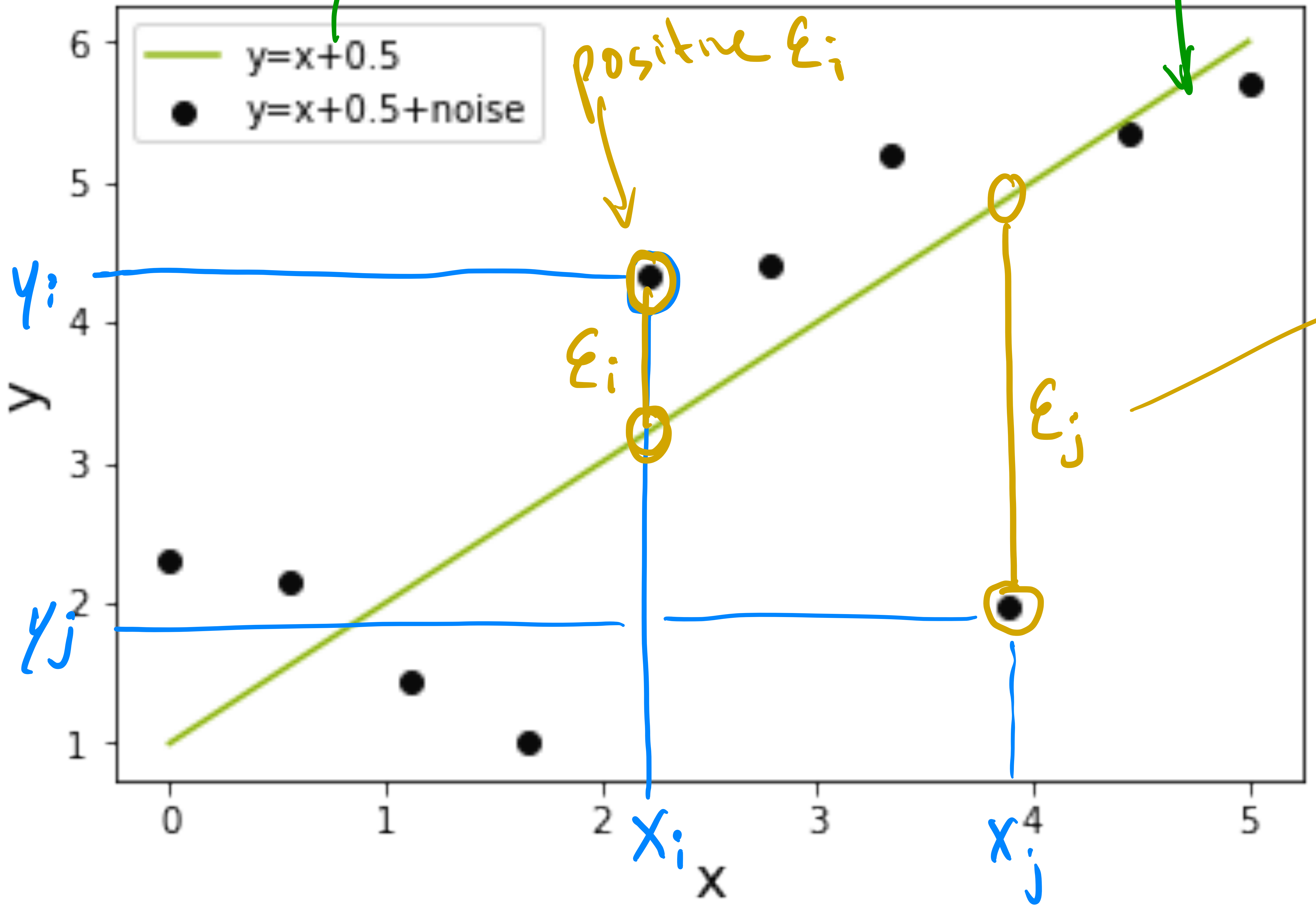
true regression line

$y = \alpha + \beta x$

- 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:

positive ε_i

negative noise ε_j



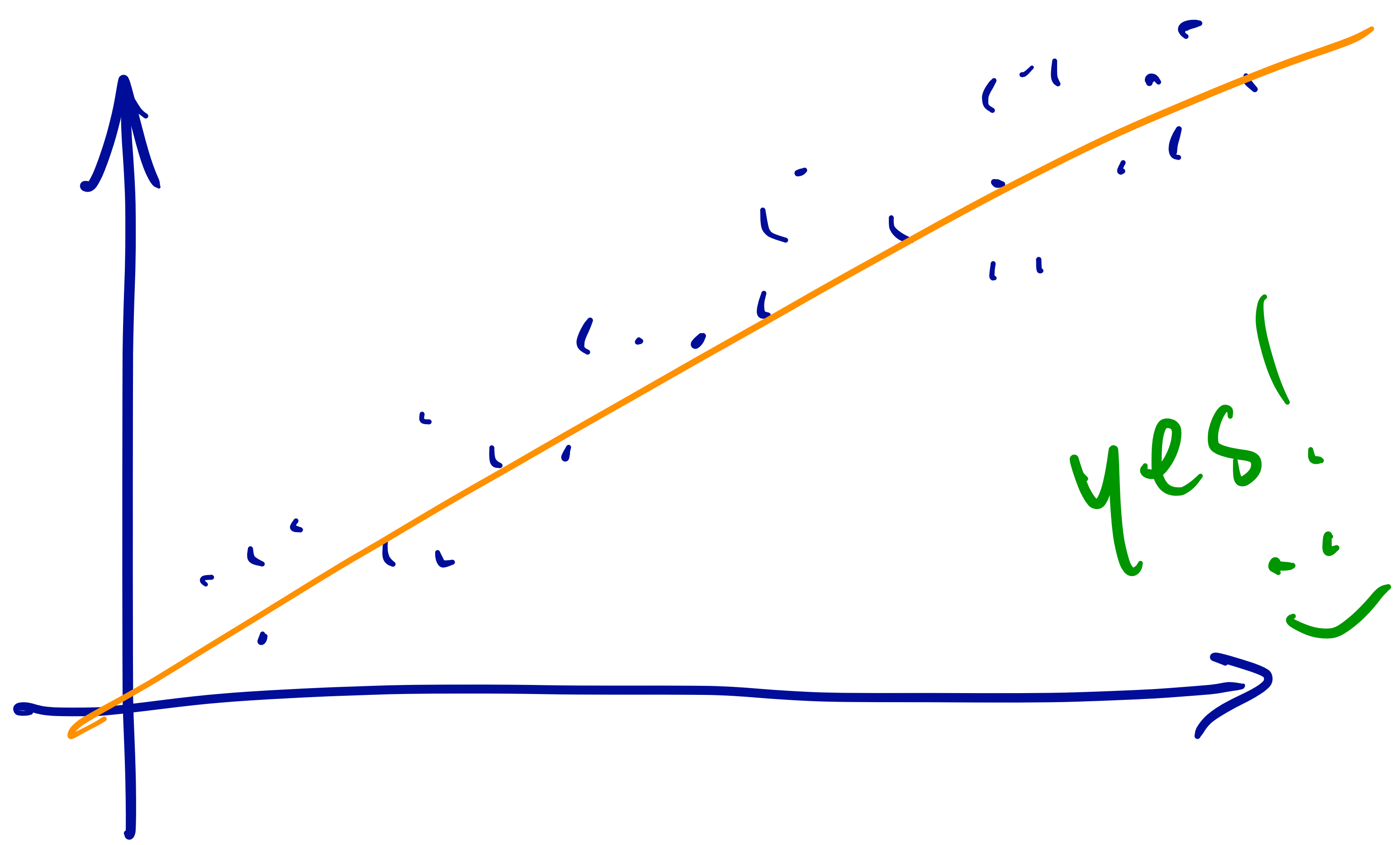
SLR: theory

- 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 between X, Y.

- Scatterplots
- Knowledge of the process generating the data.



SLR Model

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

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

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

$= \alpha + \beta X + 0$

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

- **Interpreting parameters**:

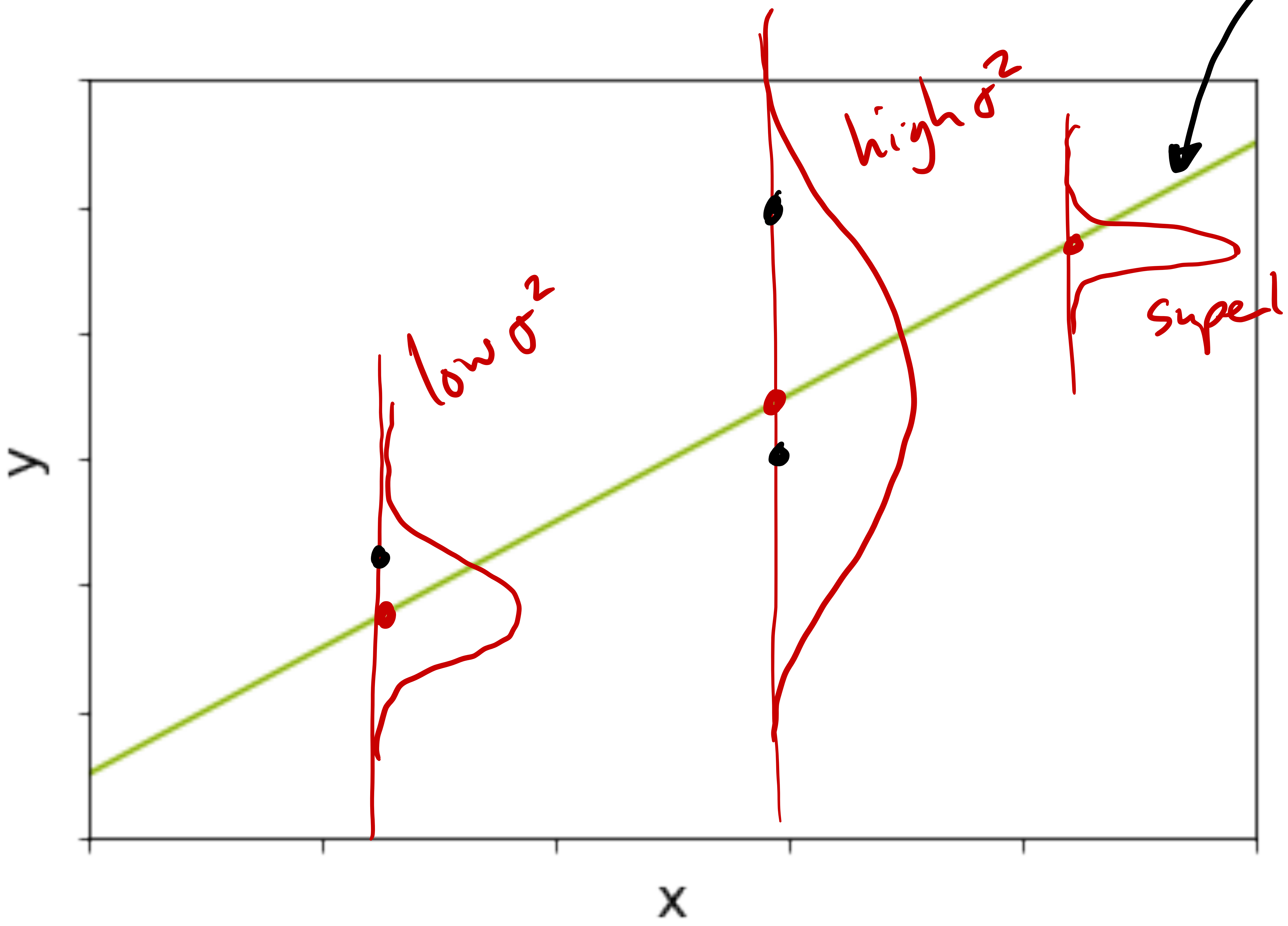
- Y is a random variable. What is its expectation, $E[Y]$? $= \alpha + \beta X$
- α (the intercept of the true regression line):
  - The average value of Y when x is zero. This is sometimes called the **baseline average**.
- β (the slope of the true regression line):
  - The average change in Y associated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

The Error Term

$y_i = \alpha + \beta x_i$

true regression line

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



- 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.

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

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

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

- 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

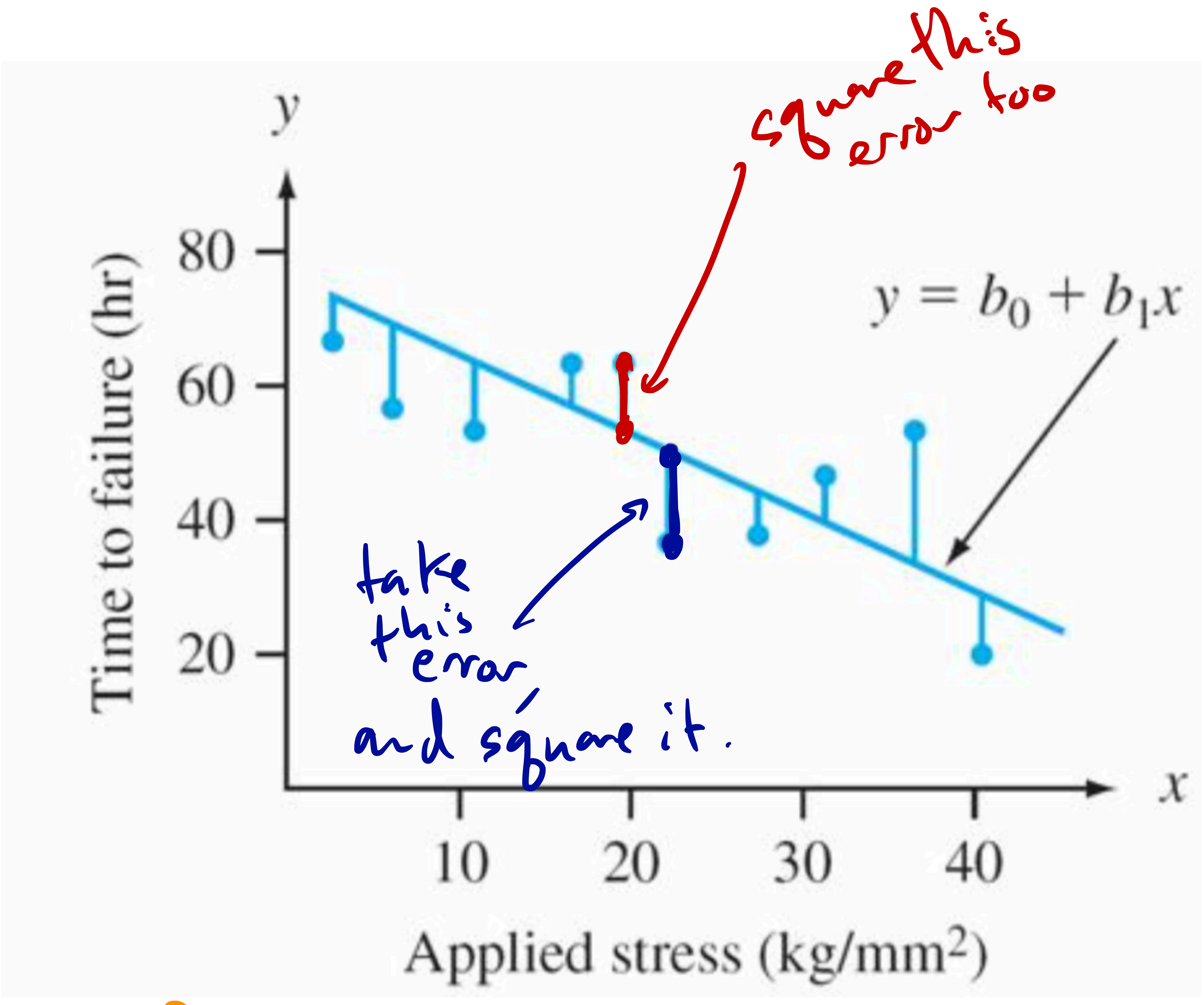
goal : cook up α , β

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

Independence 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.



in fact, square all the errors. add 'em up!

Estimating model parameters

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

$$SSE(\alpha, \beta) = \sum_{i=1}^{n} \Big(\underbrace{y_i}_{\text{actual data}} - \underbrace{(\alpha + \beta x_i)}_{\text{my predictions}} \Big)^2$$

sum of squared errors

- 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:

$\backslash hat\{\backslash alpha\}$

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

hat means "best fit" or "point estimates"

- 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

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

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

set to zero to minimize

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

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

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

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

$$\bar{y} - \alpha - \beta\bar{x} = 0$$

$$\boxed{\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}}$$ ②

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