CSCI 3022

intro to data science with probability & statistics

November 12, 2018

Statistical regression
&
Inference in Regression

Arkaive

Department of Computer Science
UNIVERSITY OF COLORADO BOULDER

Dan Larremore

Stuff & Things

- **HW6** posted tonight!. Giddyup!

Last time on CSCI3022: SLR

- Given data, $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ fit a simple linear regression of the form

$$Y_i = \alpha + \beta x_i + \epsilon_i \qquad \epsilon_i \sim N(0, \sigma^2)$$

- Compute estimates of the intercept and slope parameters by minimizing:

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

- The least-squares estimates of the parameters are:

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

Residuals

predicted

- The **fitted** or **predicted** values $\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$ 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:

$$r_i = y_i - \hat{y}_i = y_i - [\hat{\alpha} + \hat{\beta} x_i]$$

true observed

predicted

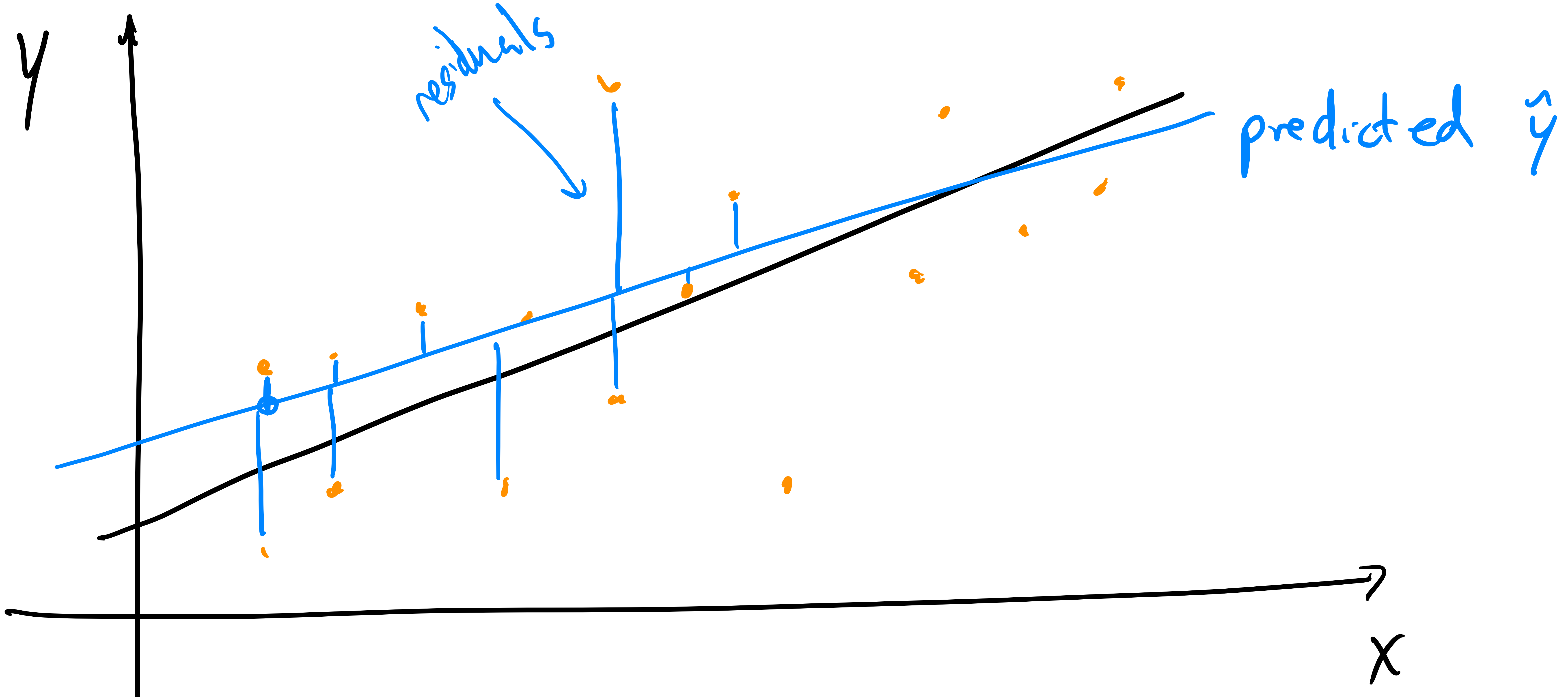
Residuals

true line

data

"best fit" regression line

- Why are the residuals estimates of the error?



Our goal was to estimate true line from data.

$\Rightarrow$ minimize SSE = minimize sum squared r_i

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

Likelihood $= P(5H, 1T \mid p) = \binom{6}{5} p^5 (1-p)^1$

- Three steps: "how likely would my data be, given p?"
    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*. $\log P(5H\,1T \mid p) = \log\binom{6}{5} + 5\log p + \log(1-p)$
    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.

$$\frac{d \log P(5H\,1T \mid p)}{dp} = 0 + \frac{5}{p} + \frac{1}{1-p}(-1) = 0$$

$$\frac{5}{p} - \frac{1}{1-p} = 0$$

$$\frac{5}{p} = \frac{1}{1-p}$$

$$5(1-p) = p$$

$$5 - 5p = p$$

$$5 = 6p$$

$$\text{so} \quad \boxed{\hat{p} = \frac{5}{6}}$$

Maximum likelihood estimates

MLE (generally)

- **Maximum Likelihood Estimation** asks: what are the *parameters* that best explain the data that we see?
- **In practice**, this means that we usually go through three steps:

1. Write down the probability of getting the data, given the probability distribution and the parameter(s) of interest. (This is the likelihood.)
2. Take a log to get the *log-likelihood*.
3. Take a derivate with respect to the parameter, set equal to zero, and solve to find the MLE value of the parameter. (Don’t forget to put a hat on it 🎩)

MLE for simple linear regression

normal distrib. PDF

$$Y_i = \alpha + \beta x_i + N(0, \sigma^2)$$

$$= N(\alpha + \beta x_i, \sigma^2)$$

$$\log(abc) = \log a + \log b + \log c$$

1. $$P(\text{data} \mid \alpha, \beta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - (\alpha + \beta x_i))^2}{2\sigma^2}}$$

2. $$\text{log Likelihood} = \sum_{i=1}^{n} \log \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{(y_i - (\alpha + \beta x_i))^2}{2\sigma^2}$$

$$= n \log \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2\sigma^2} \underbrace{\sum_{i=1}^{n} (y_i - (\alpha + \beta x_i))^2}_{SSE}$$

3. Minimizing SSE $\equiv$ Maximizing Likelihood!

1. P(data | params)
2. Take a log.
3. Derivative = 0

The punchline:

- Maximum Likelihood and Least-Squares are solving **the same problem**
- Important: this means that when we are solving the least-squares problem, what are we *always, implicitly assuming about the errors?*

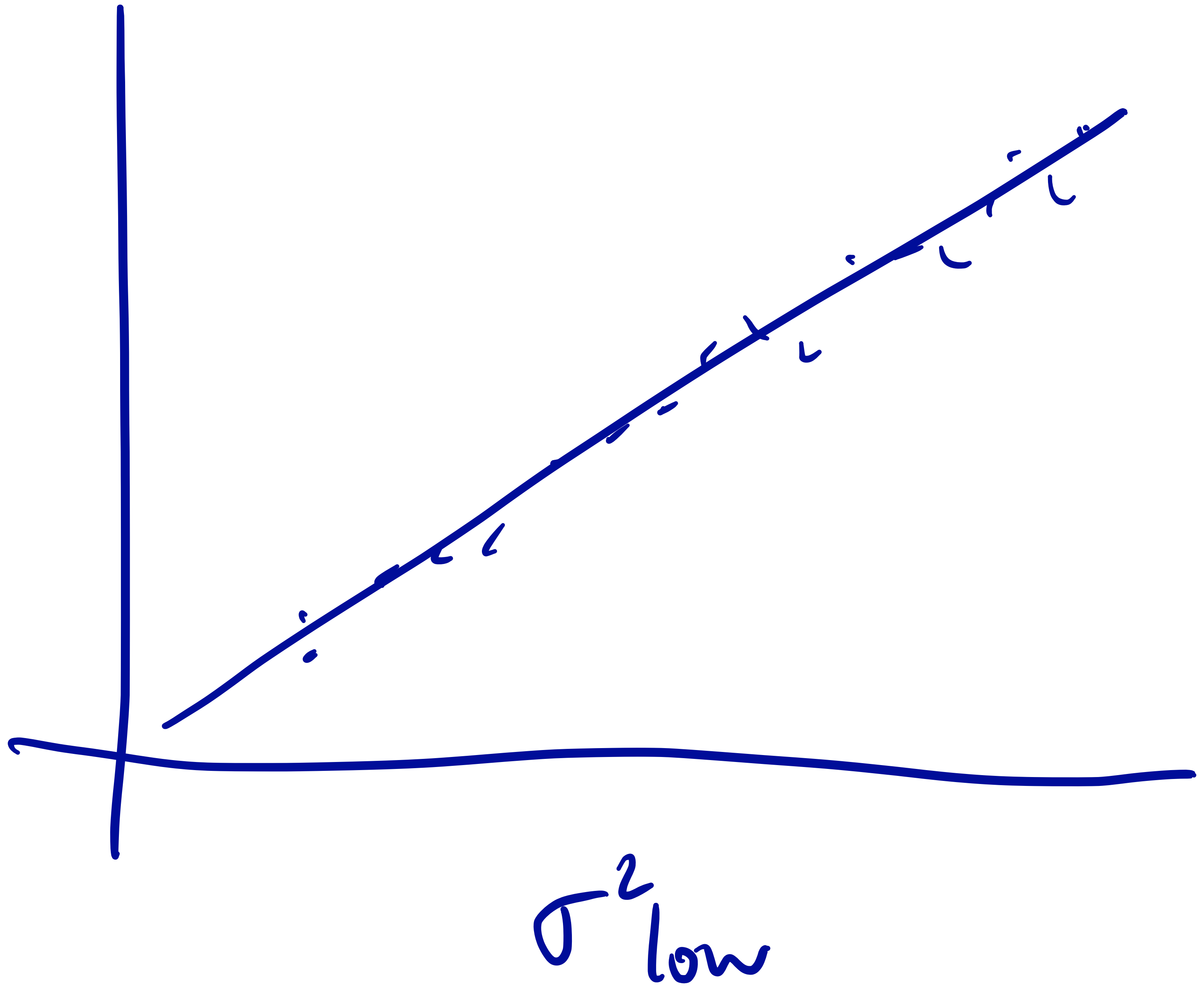
- errors are added to y_i $\qquad$ $y_i = \alpha + \beta x_i \underline{+ \varepsilon_i}$
- each ε_i is indep of others.
- $\varepsilon_i \sim N(0, \sigma^2)$

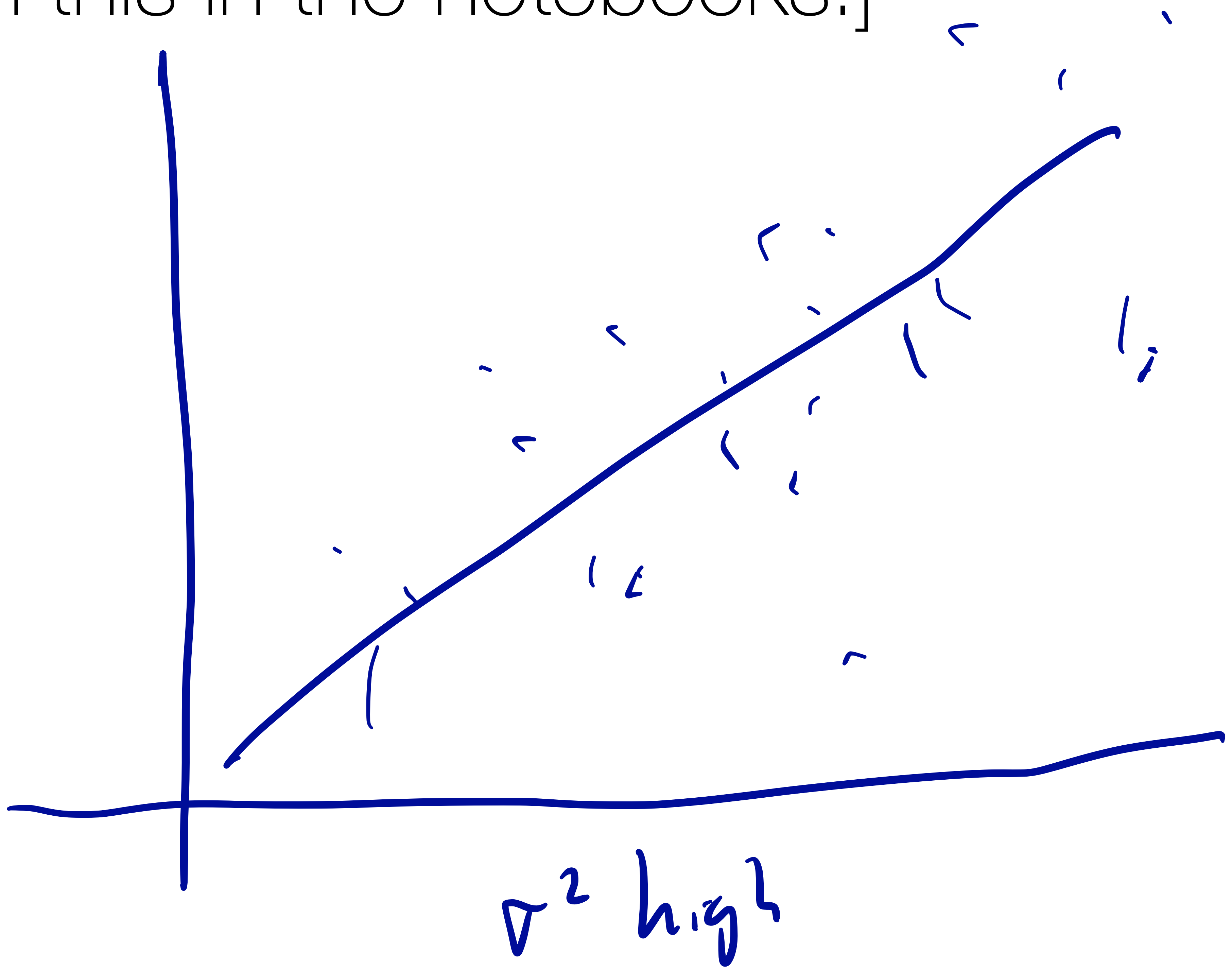
For the rest of today:

- **How can we**:
  - Estimate the variance in the population of estimates?
  - Quantify the goodness-of-fit in our simple linear regression model?
  - Perform inference on the regression parameters?

Estimating the variance

- The parameter σ^2 determines the spread of the data about the true regression line. [We experimented with this in the notebooks!]





generally, we don't know what σ^2 is!

Estimating the variance

$$\boxed{\hat{\sigma}^2 = \frac{SSE}{n-2}}$$

- The divisor (n-2) in the estimate of σ^2 is the number of *degrees of freedom* (abbreviated df) associated with the estimate of SSE.

- This is because to obtain $\hat{\sigma}^2$, the two parameters $\hat{\alpha}$ and $\hat{\beta}$ must first be estimated, which results in a loss of 2 degrees of freedom.

<u>note:</u> $\bar{x} = \frac{1}{n} \sum_i x_i$

$$s^2 = \frac{1}{n-1} \sum_i (x_i - \bar{x})^2$$

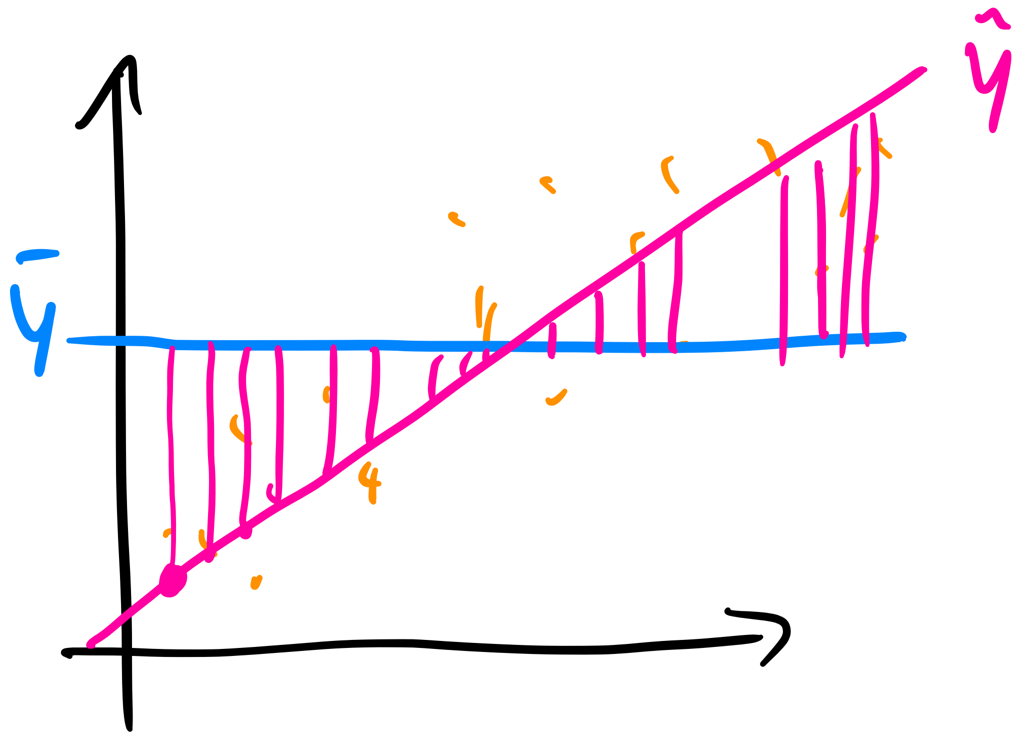
↑ loss of one d.o.f.

The coefficient of determination

- The coefficient of determination, R^2 quantifies how well the model explains the data.

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \leftarrow$$

how much uncertainty / variance in y_i remains unexplained after fitting model $\hat{y}_i$.

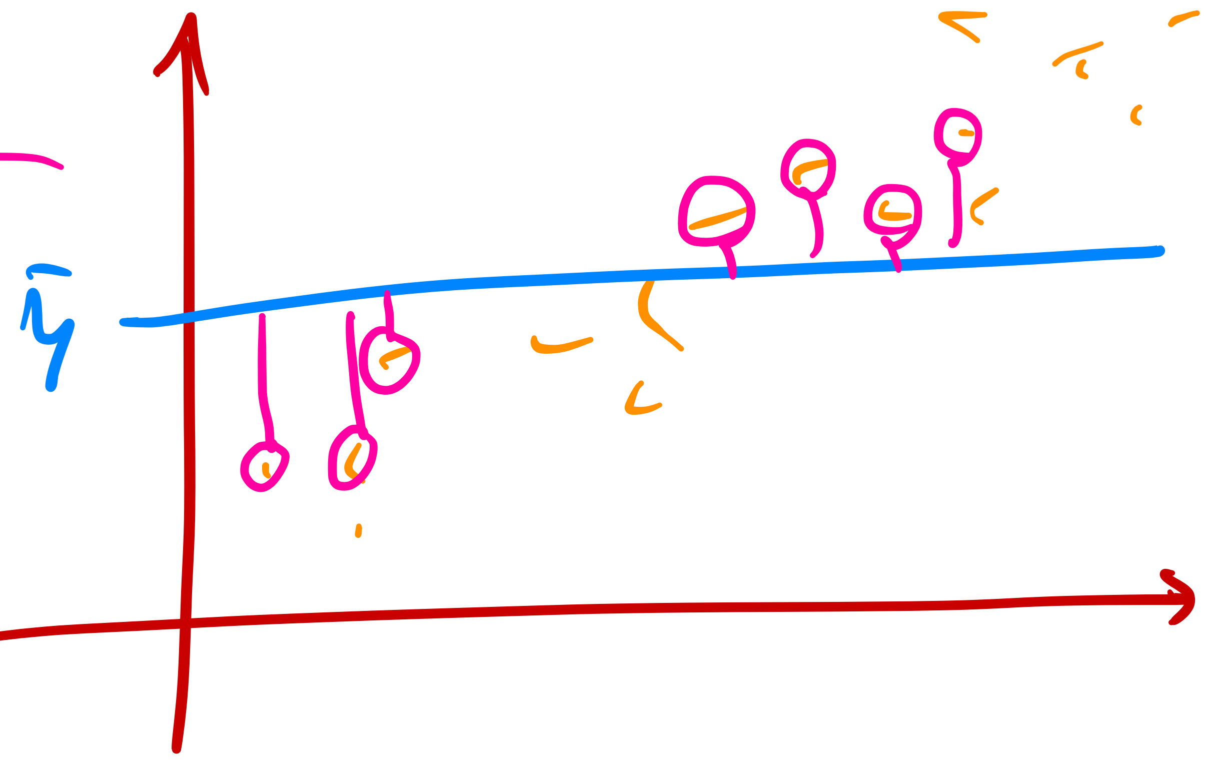


$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \leftarrow$$

regression sum of squares

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2 \leftarrow$$

total sum of squares.



- R^2 is a value between 0 and 1.

$$SST = SSR + SSE$$

= what can be explained by regression + what can't be explained by regression

$$\boxed{R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}} \leq 1$$

The coefficient of determination

The **sum of squared errors** (SSE)

see prev slide

can be interpreted as a measure of how much variation in *y* is left unexplained by the model: how much variation *cannot* be attributed to a linear relationship?

The **regression sum of squares** is given by

see prev. slide

A quantitative measure of the total amount of variation in observed *y* values is given by the so-called **total sum of squares**

SST see prev. slide

The coefficient of determination

- The sum of squared deviations about the least-squares line is smaller than the sum of squared deviations about any other line, i.e. $\boxed{SSE < SST}$ unless the horizontal line itself is the least-squares line

- The ratio SSE/SST is the proportion of total variation in the data that cannot be explained by the simple linear regression model, and the coefficient of determination is

$$SST = \sum (y_i - \bar{y})^2$$

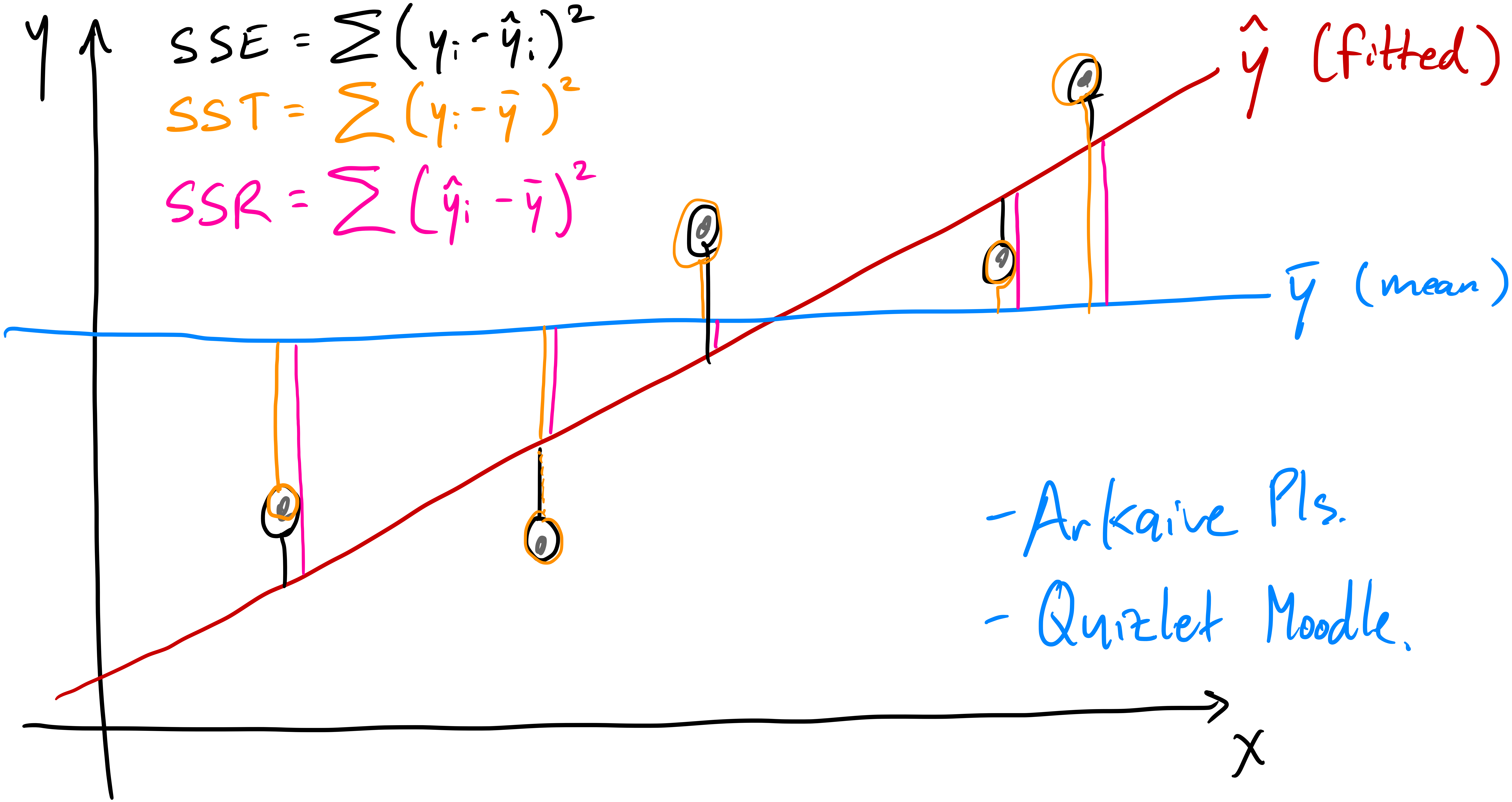
$$SSE = \sum (y_i - \hat{y}_i)^2$$

$$R^2 = 1 - \frac{SSE}{SST}$$

variance <u>after</u> modeling

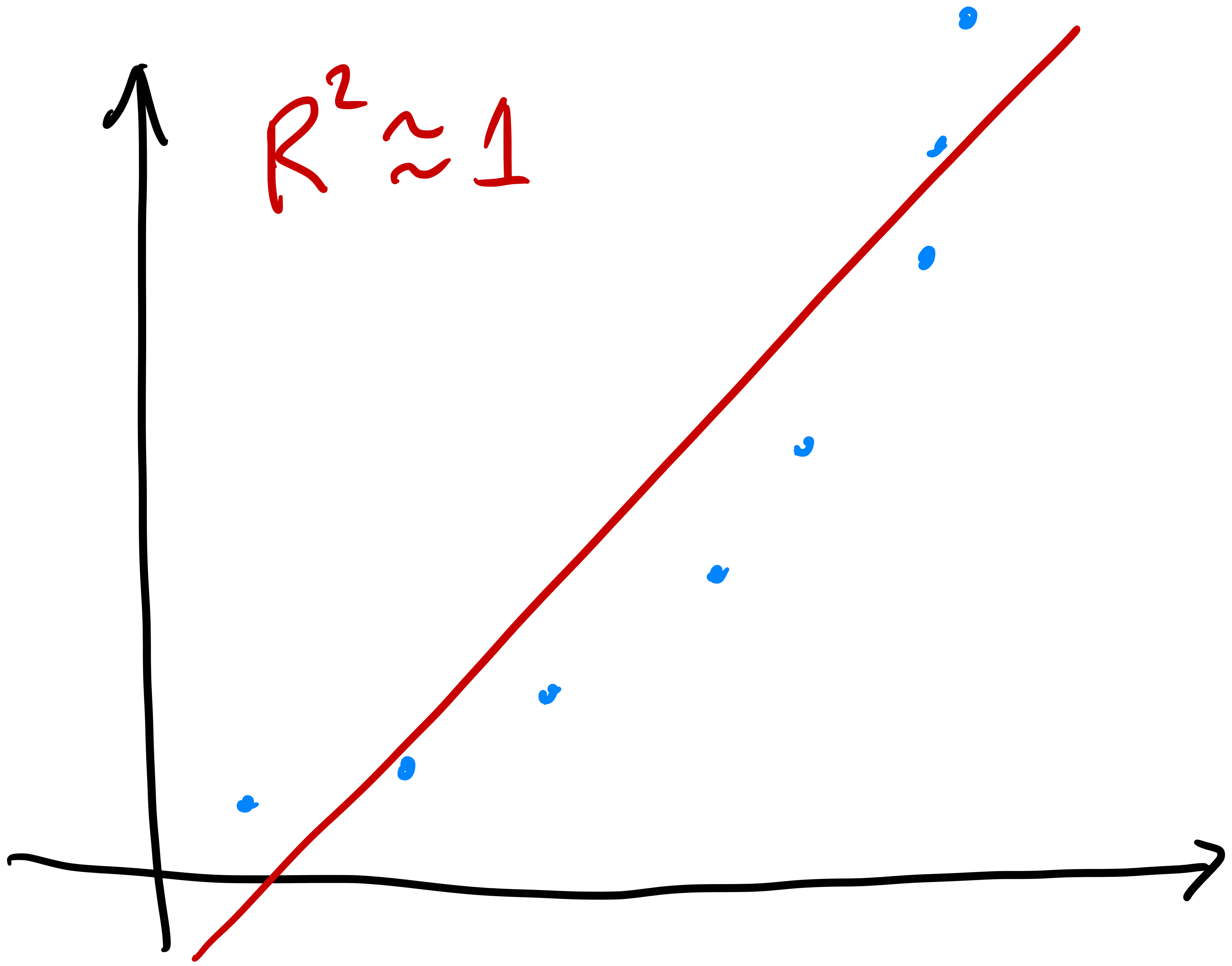
variance <u>before</u> modeling.

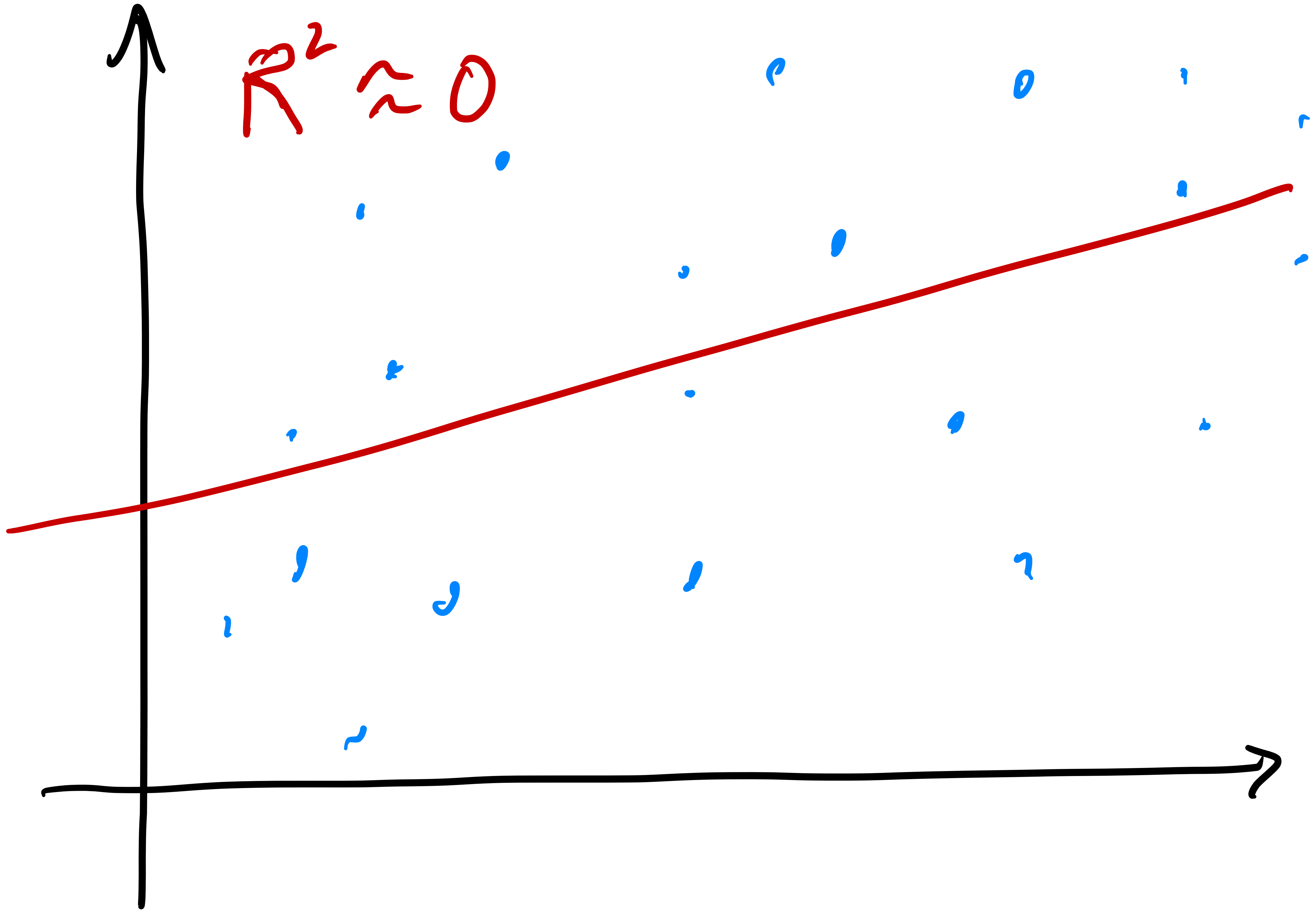
The coefficient of determination



The coefficient of determination

- Note: R^2 is the proportion of total variation in the data that is explained by the model.
- But: R^2 does *not* tell you that you necessarily have the correct model!





Inference about parameters

- The parameters in simple linear regression have distributions! We demonstrated this in the in-class notebook last time.
- From these distributions, we can conduct hypothesis tests (e.g.: $\beta = 0$), compute confidence intervals, etc.

  $\beta = 5$
- **Distributions**:

$H_0 : \beta = c$

$H_1 : \beta \neq c$

Specific: $H_0 : \beta = 0$ (no trend -or- y does not depend on x)

$H_1 : \beta \neq 0$

$$\hat{\beta} \sim N\left(\beta, \frac{\sigma^2}{\sum_i (x_i - \bar{x})^2}\right) = N\left(\beta, \frac{\frac{SSE}{n-2}}{\sum_i (x_i - \bar{x})^2}\right)$$

Inferences about the parameters

- **Confidence intervals**: $100(1-\alpha)\%$ CI is given by:

$$\hat{\beta} \pm t_{\alpha/2,\, n-2} \times SE(\hat{\beta}) = \boxed{\hat{\beta} \pm t_{\alpha/2,\, n-2} \underbrace{\sqrt{\frac{\frac{SSE}{n-2}}{\sum_i (x_i - \bar{x})^2}}}}$$

d.o.f.

- **Tests**:

$H_0: \beta = 0$

$H_1: \beta \neq 0$

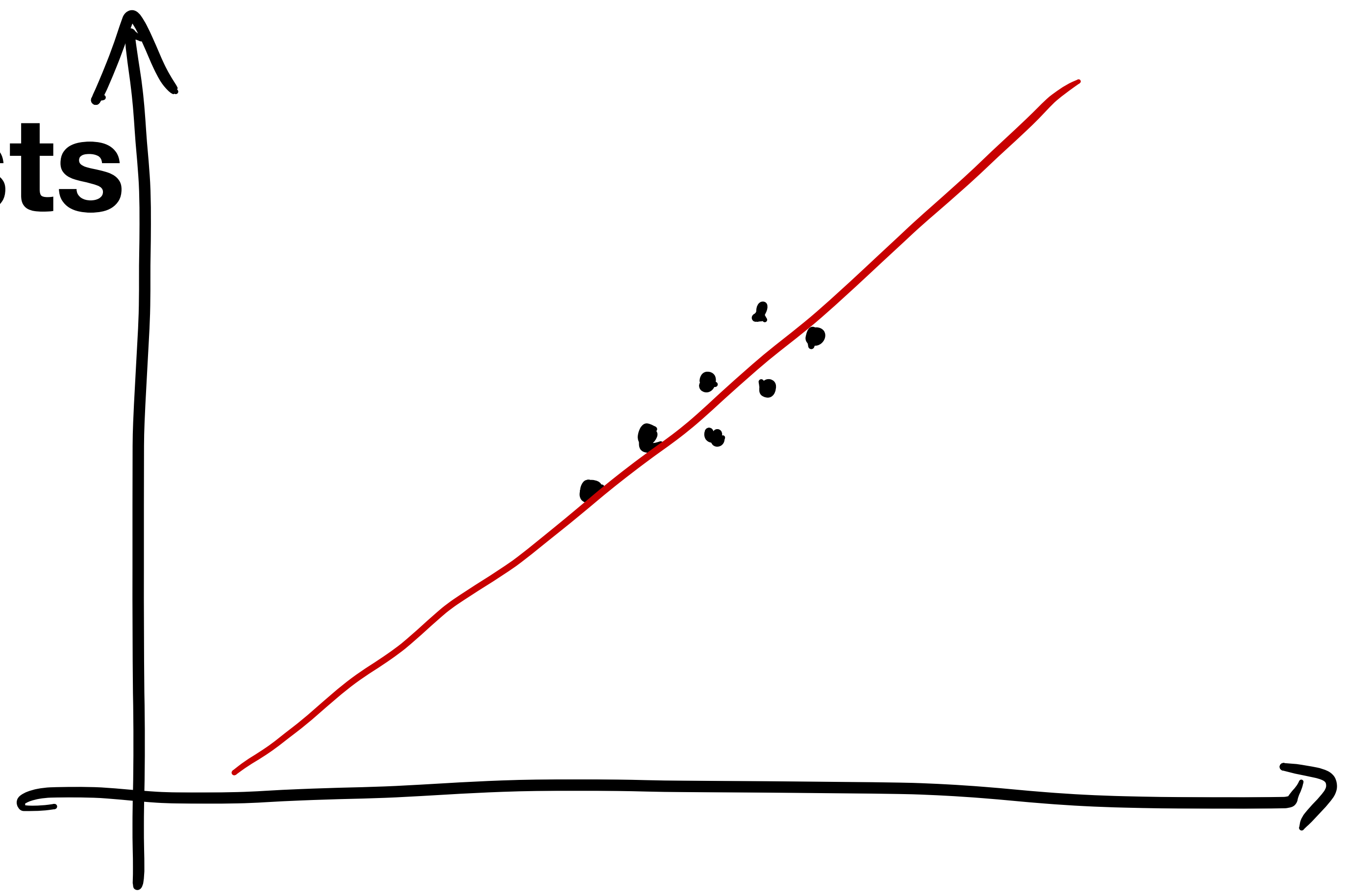
test statistic $t = \dfrac{\hat{\beta} - 0}{SE(\hat{\beta})}$

Inferences about the parameters

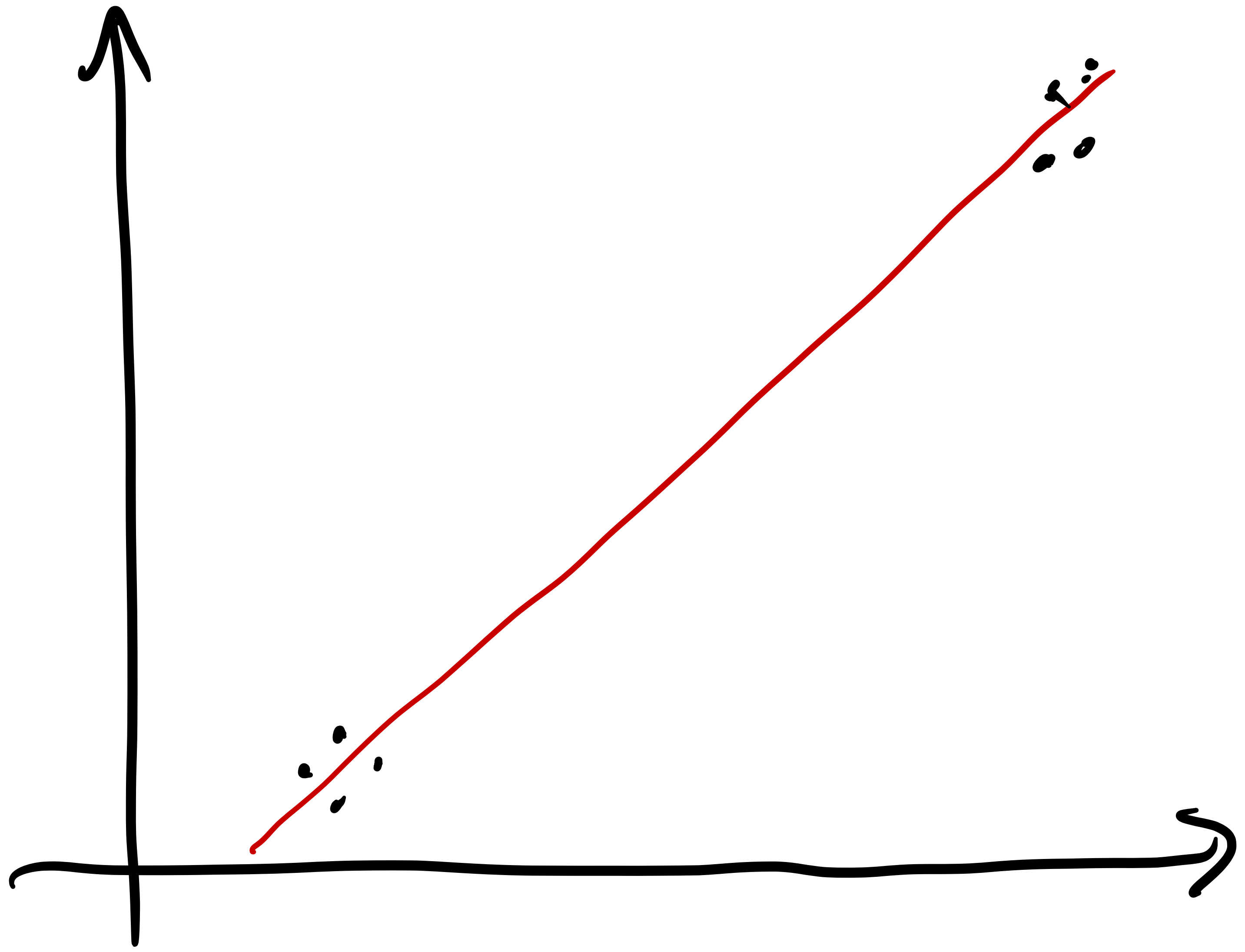
- **Confidence intervals**:

What does $\sum_i (x_i - \bar{x})^2$ mean?

- **Tests**



small



large