CSCI 3022

intro to data science with probability & statistics

Lecture 23
April 9, 2018

Statistical regression
and
Inference in Regression

Department of Computer Science
UNIVERSITY OF COLORADO BOULDER

Tony
~~Dan Larremore~~

Stuff & Things

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

Last time on CSCI3022: SLR

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

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

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

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

- The least-squares estimates of the parameters are:

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

Does it work?

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

Residuals

Fixed

- 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:

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

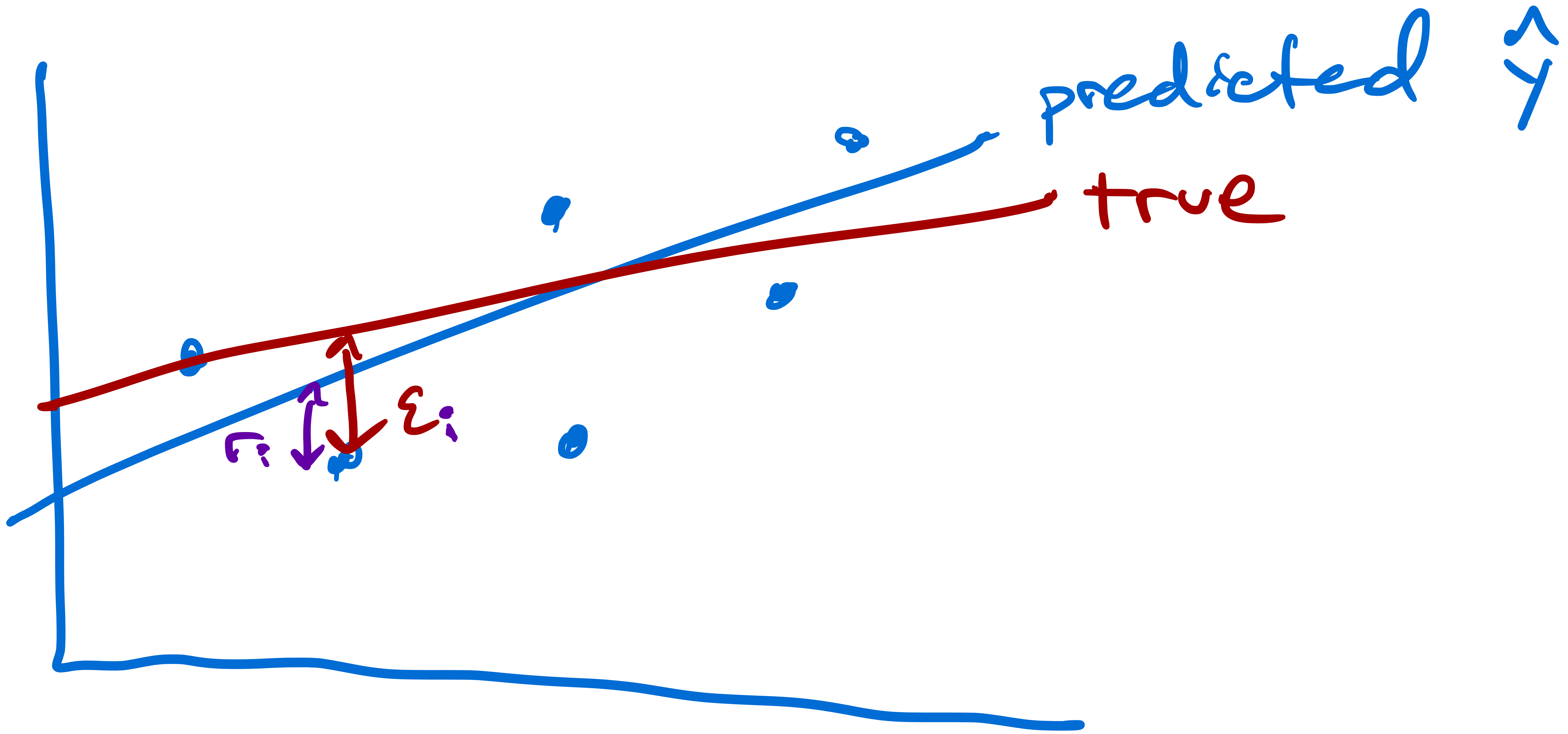
Residuals

truth
$$y = \alpha + \beta x$$

$N(0, \sigma^2)$

measure: $y_i = \alpha + \beta x_i + \varepsilon_i$

- Why are the residuals estimates of the error?



Want to estimate true line as well as possible

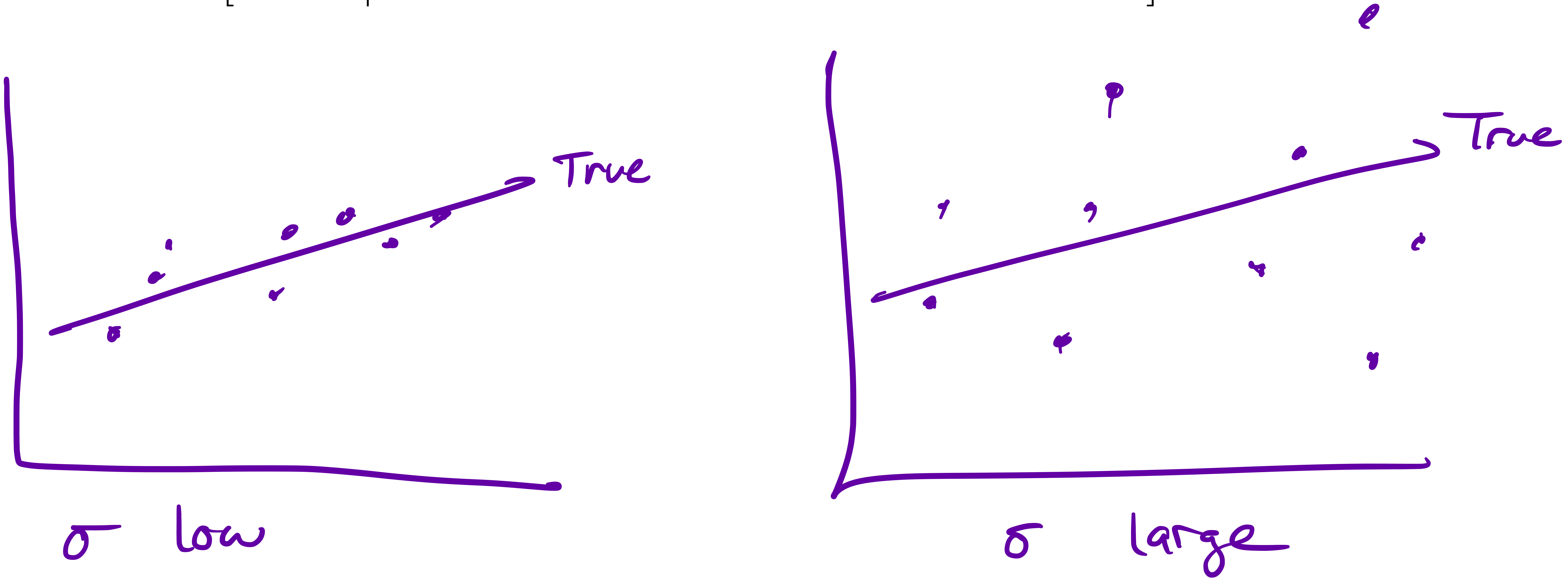
$\rightarrow$ minimize sum of squared r_i (SSE)

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 is!

Estimating the variance

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

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

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

- 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.

Mean: $\bar{x} = \frac{1}{n} \sum x_i$

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

The coefficient of determination

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

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

how much uncertainty variance in y_i can be explained by model $\hat{y}_i$

$$SSR = \sum (\hat{y}_i - \bar{y})^2$$

regression sum of squares

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

$SST = SSR + SSE =$ what can be explained by regression + what can't be explained by regression

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

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

The coefficient of determination

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

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

SSR = defined before

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

SST

The coefficient of determination

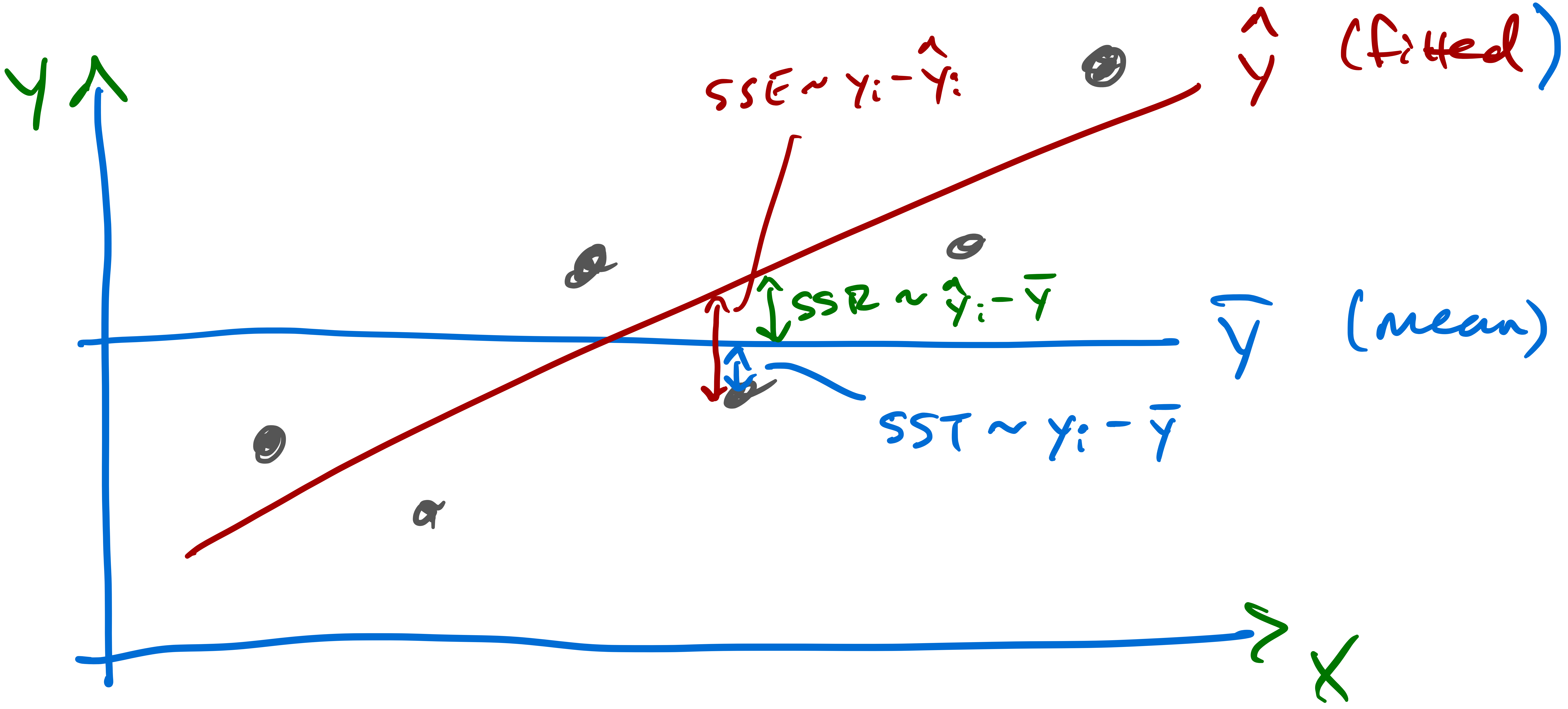
- The sum of squared deviations about the least-squares line is smaller than the sum of squared deviations about any other line, i.e. SSE < SST unless the horizontal line itself is the least-squares line
- The ratio SSE/SST is the proportion of total variation in the data that cannot be explained by the simple linear regression model, and the coefficient of determination is

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

variability we can't explain w/ SLR

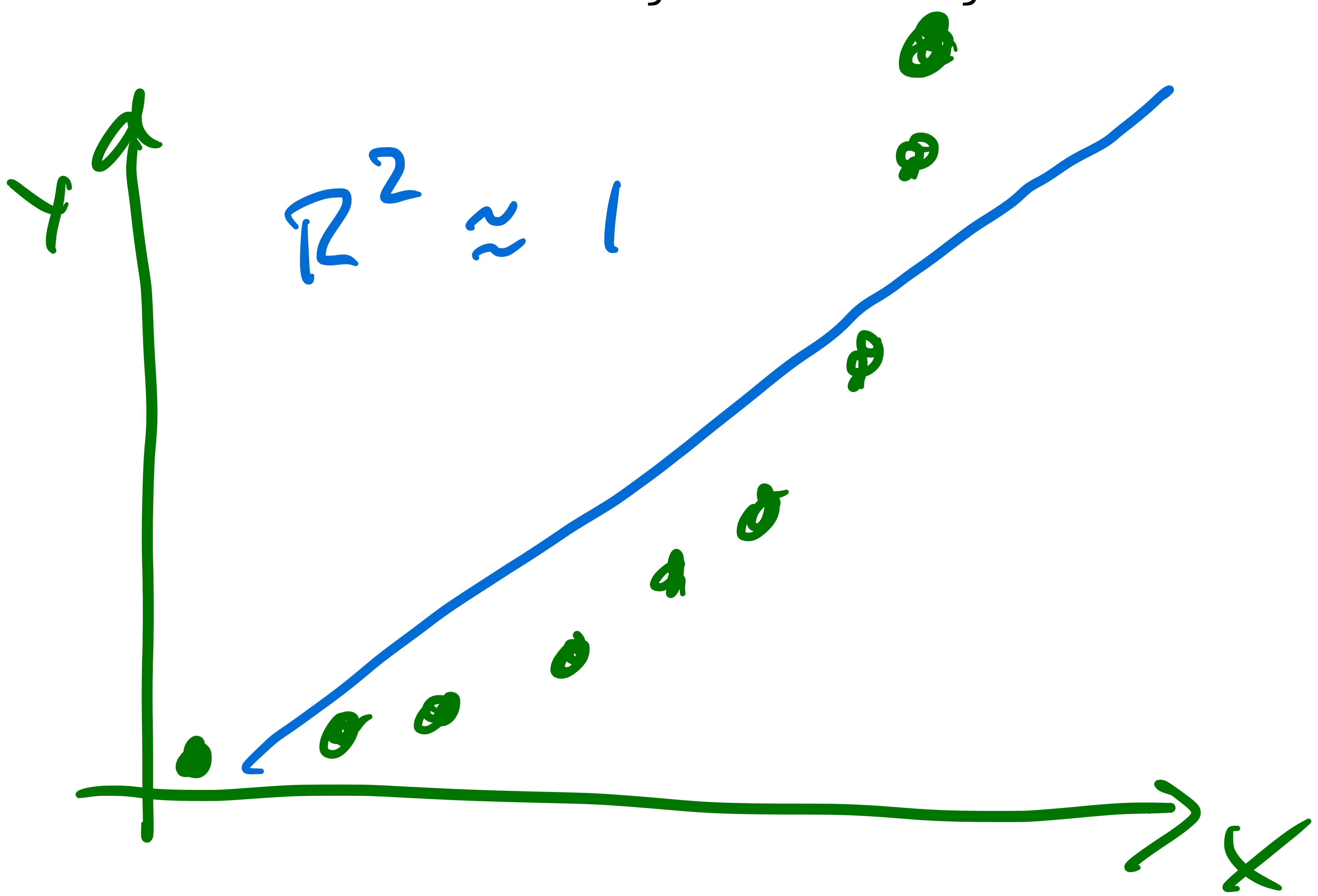
total variability

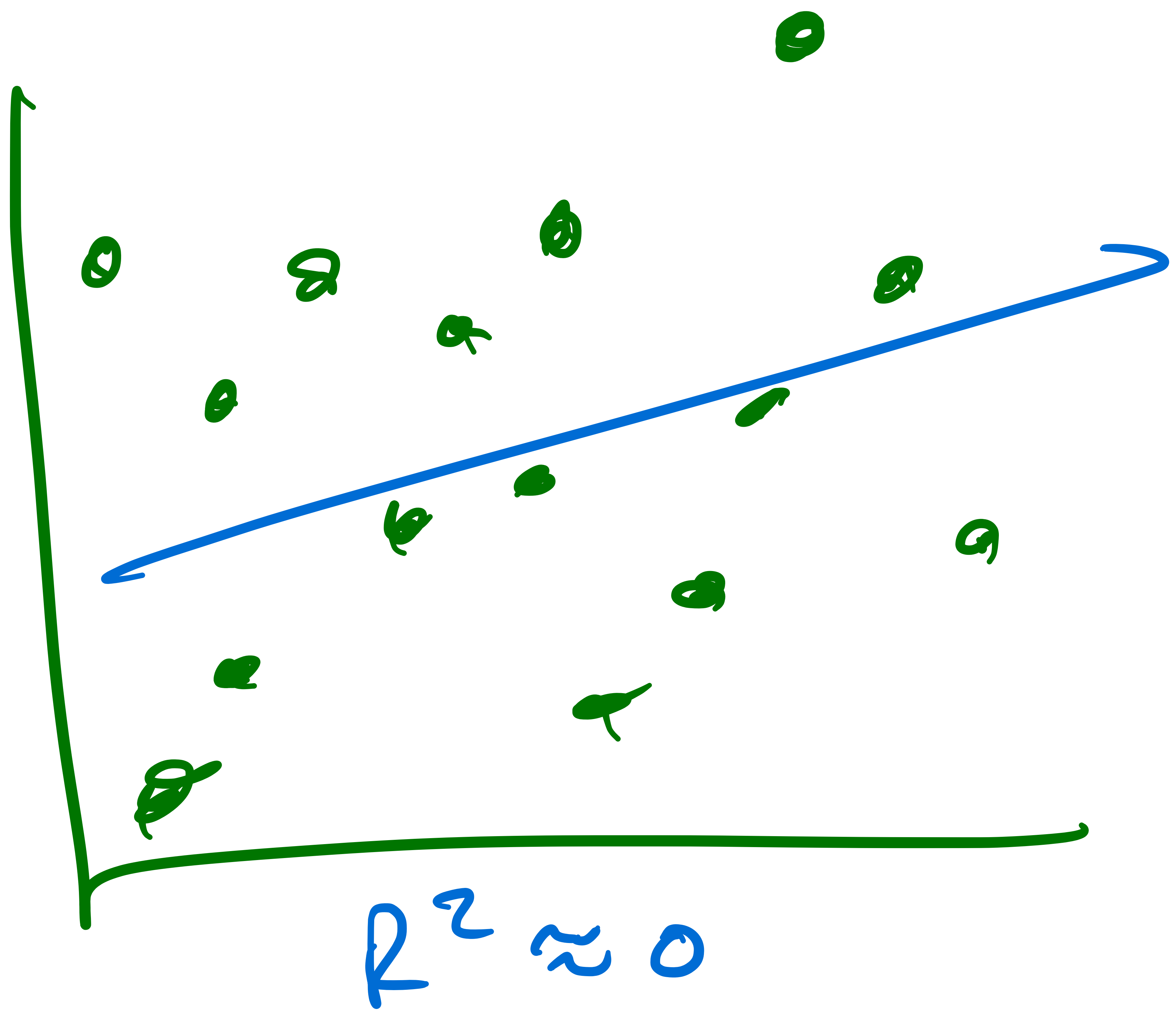
The coefficient of determination



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.: H?), compute confidence intervals, etc.
- **Distributions**: especially for β : $H_0: \beta = c$ (e.g. 0) $H_1: \beta \neq c$

$$\hat{\beta} \sim N\Big(\beta, \underbrace{\frac{\sigma^2}{\sum_i (x_i - \bar{x})^2}}\Big)$$

$$SE(\hat{\beta}) = \frac{SSE}{n-2} \Big/ \sum_i (x_i - \bar{x})^2$$

Inferences about the parameters

- **Confidence intervals**: $100 \times (1-\alpha)\,\%$ CI :

$$\hat{\beta} \pm t_{\alpha/2,\, n-2} \times SE(\hat{\beta})$$

- **Tests**:

$H_0: \beta = 0$? $\longrightarrow$ Test statistic: $t = \dfrac{\hat{\beta} - 0}{SE(\hat{\beta})}$ $\longrightarrow$ Compute p-value OR C.I.

$H_1: \beta \neq 0$