Welcome to STA 101!

Statistics is a confrontation with uncertainty.

Statistics confronts uncertainty by quantifying it.

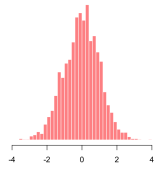
Data analysis

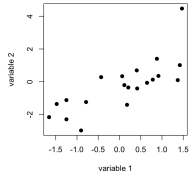
Transforming messy, incomplete, imperfect data into knowledge:

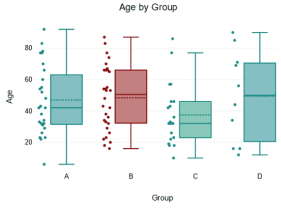
subject	variable_1	variable_2
1	-1.65692830	-2.16524631
2	-0.90396488	-2.97993045
3	1.37141732	0.09720280
4	-0.43176527	0.27970110
5	0.40649190	0.69143221
6	1.47092198	4.47233461
7	-0.78625051	-1.24276055
8	0.64835135	-0.06749005
9	0.06363568	0.33517580

`ggplot` $\Longrightarrow$

`|>` $\Longrightarrow$







	mean	median	std
Variable 1	0.795	-0.292	0.200
Variable 2	0.343	0.616	0.834
Variable 3	1.587	0.508	2.501

Statistical inference

Quantifying our uncertainty about that knowledge:

- **Question**: What's the number?
- **Answer**: best-guess $\pm$ margin-of-error

Today's topic

Simple linear regression.

Goal

Study the association between two variables.

Subtext

Assess the causal impact of one variable on another.

Warning

Association alone does not imply causation.

Bottom line

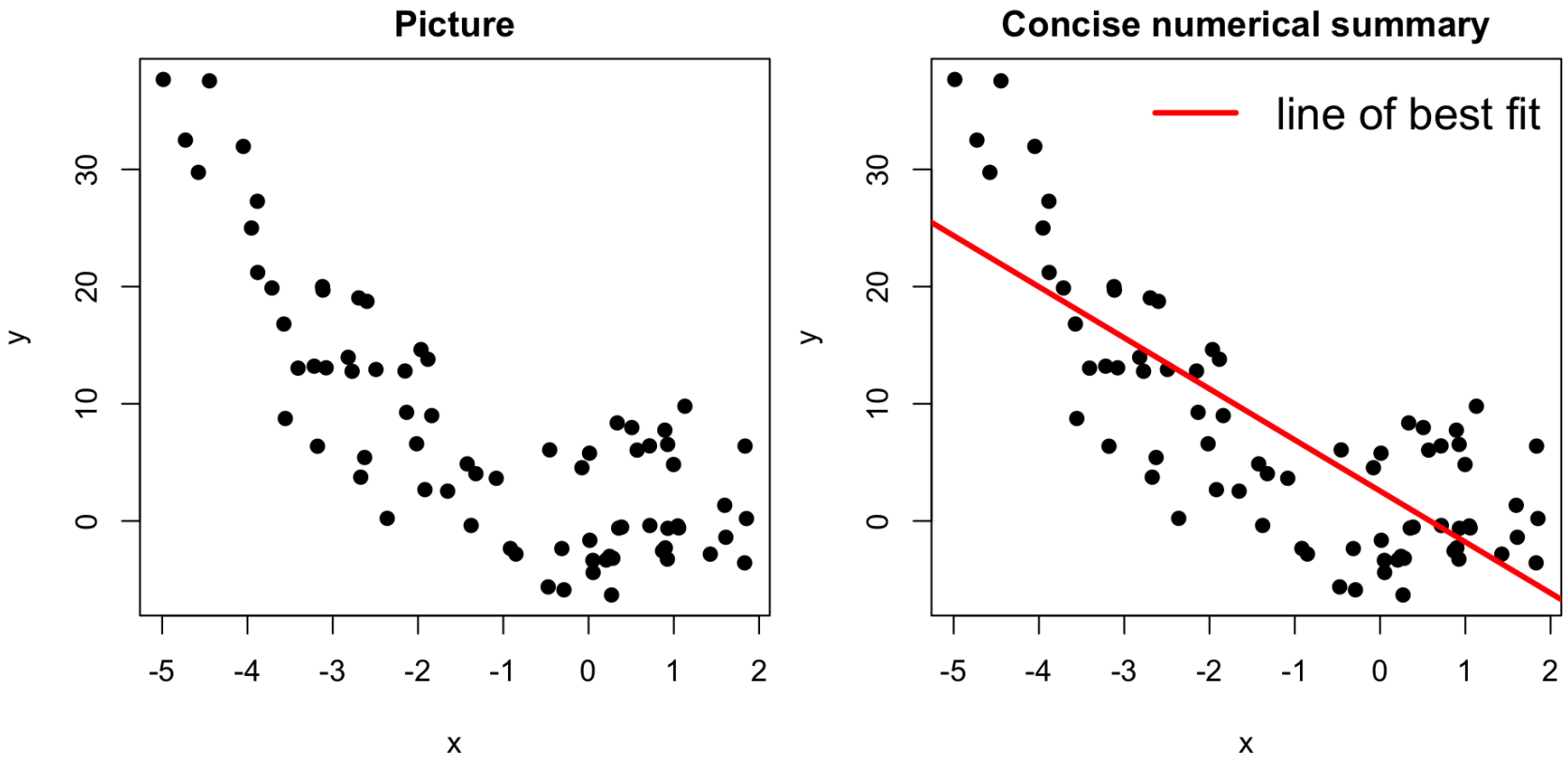
"Mere" association will still be an important ingredient.

“Association alone does not imply causation.”

- there are plenty of meaningless coincidences out there;
  - `https://tylervigen.com/spurious-correlations`
- does x cause y, or does y cause x?
- are you sure there isn't some third thing going on?

Data analysis for two variables

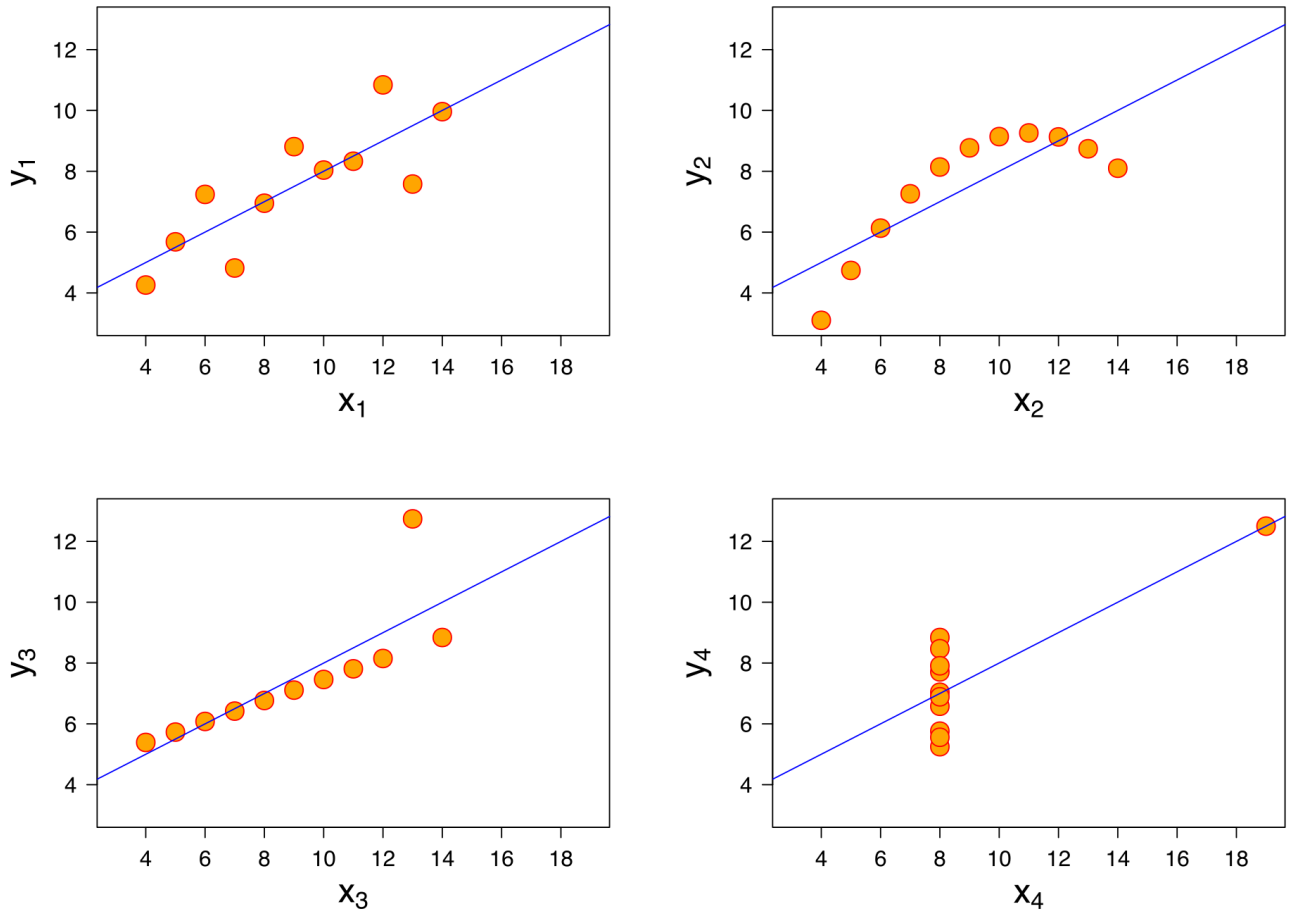
Turn messy, bivariate data into knowledge...



Theme: these need to work together!

Anscombe's quartet

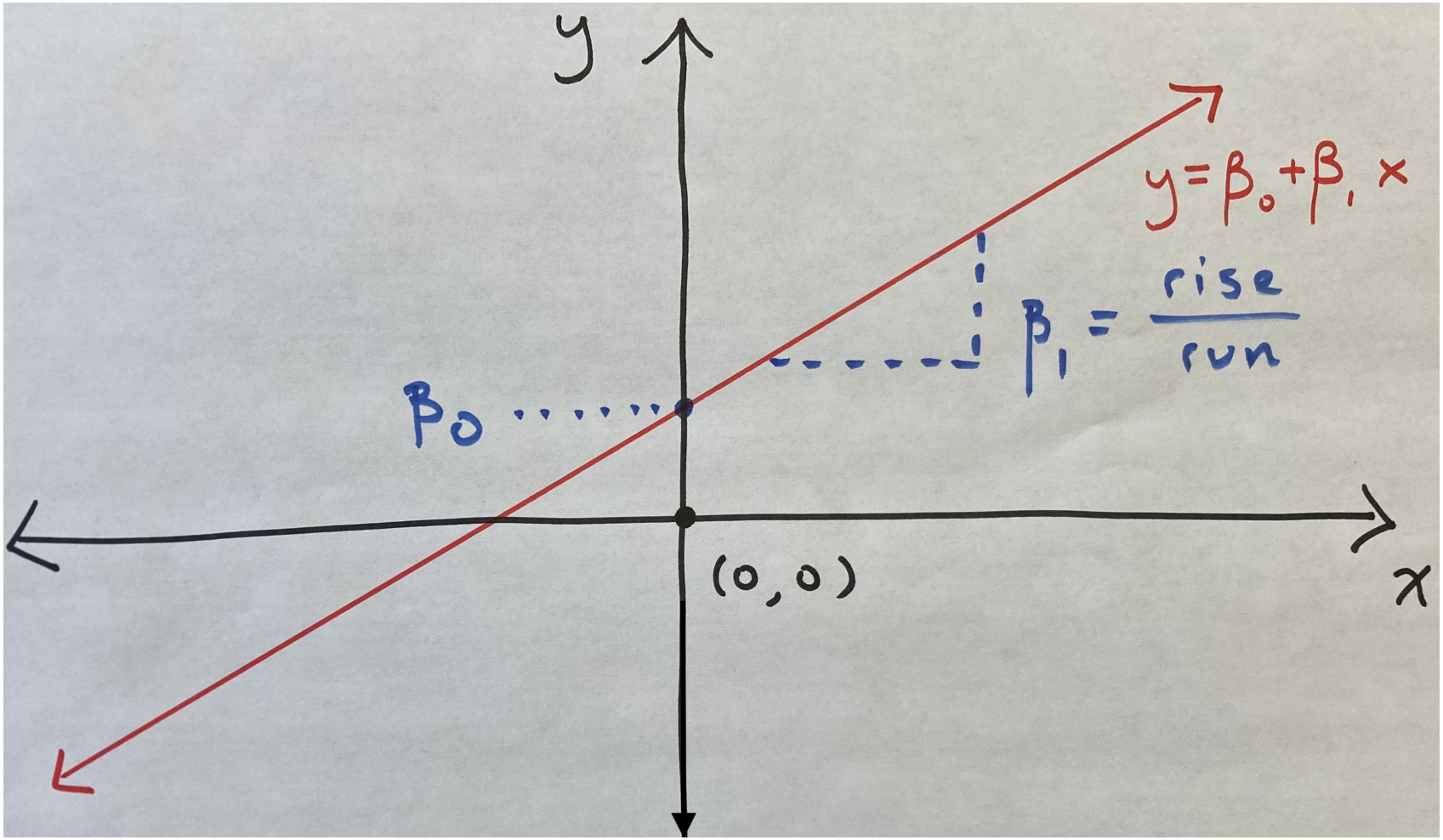
These all have the same line of best fit:



ABV: **A**lways **B**e **V**isualizing

Review: straight lines in the xy-plane

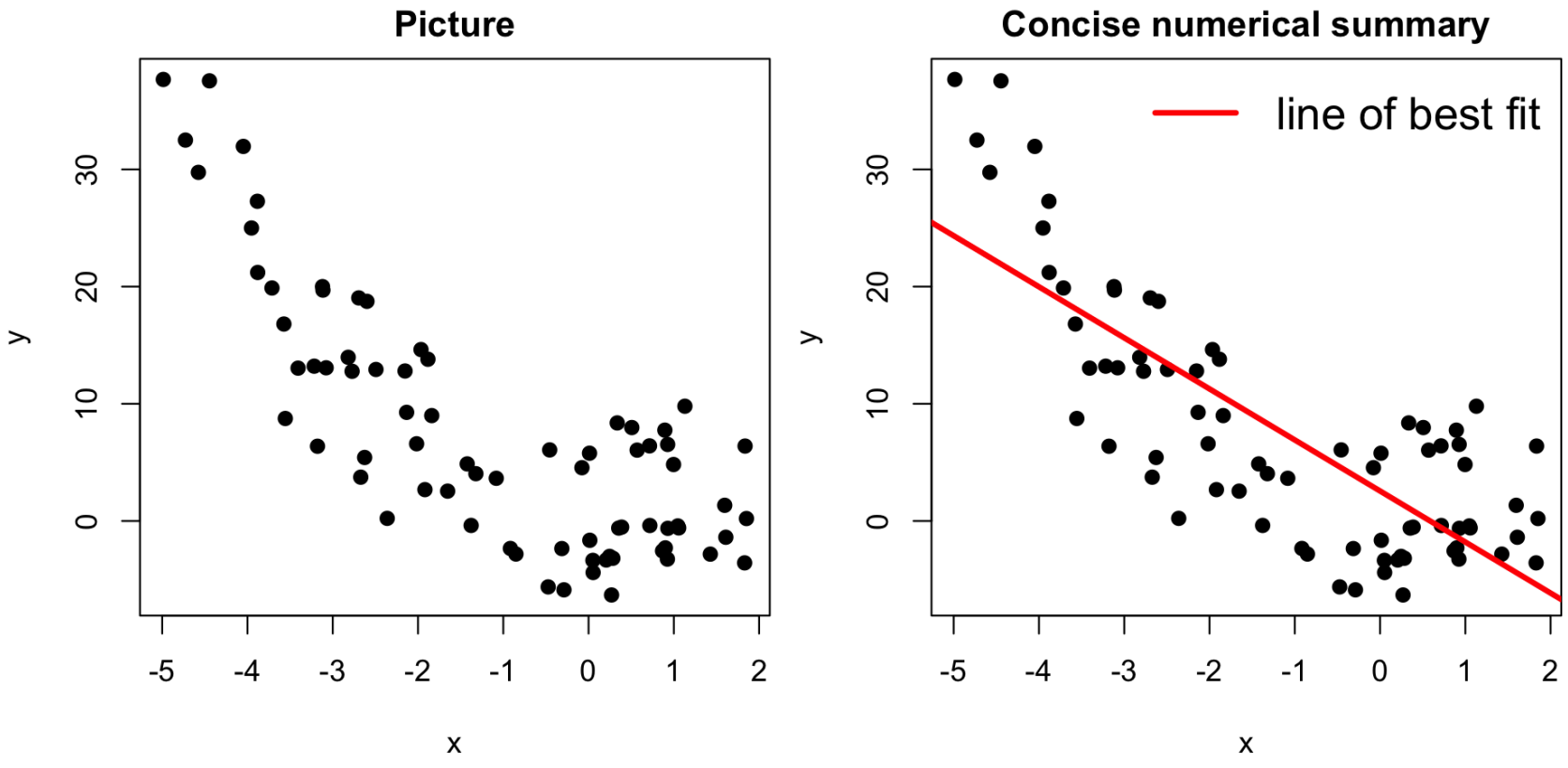
$$y = \underbrace{\beta_0}_{\text{intercept}} + \underbrace{\beta_1}_{\text{slope}} x.$$



(β is the Greek letter *beta*, pronounced "bae, duh.")

Data analysis for two variables

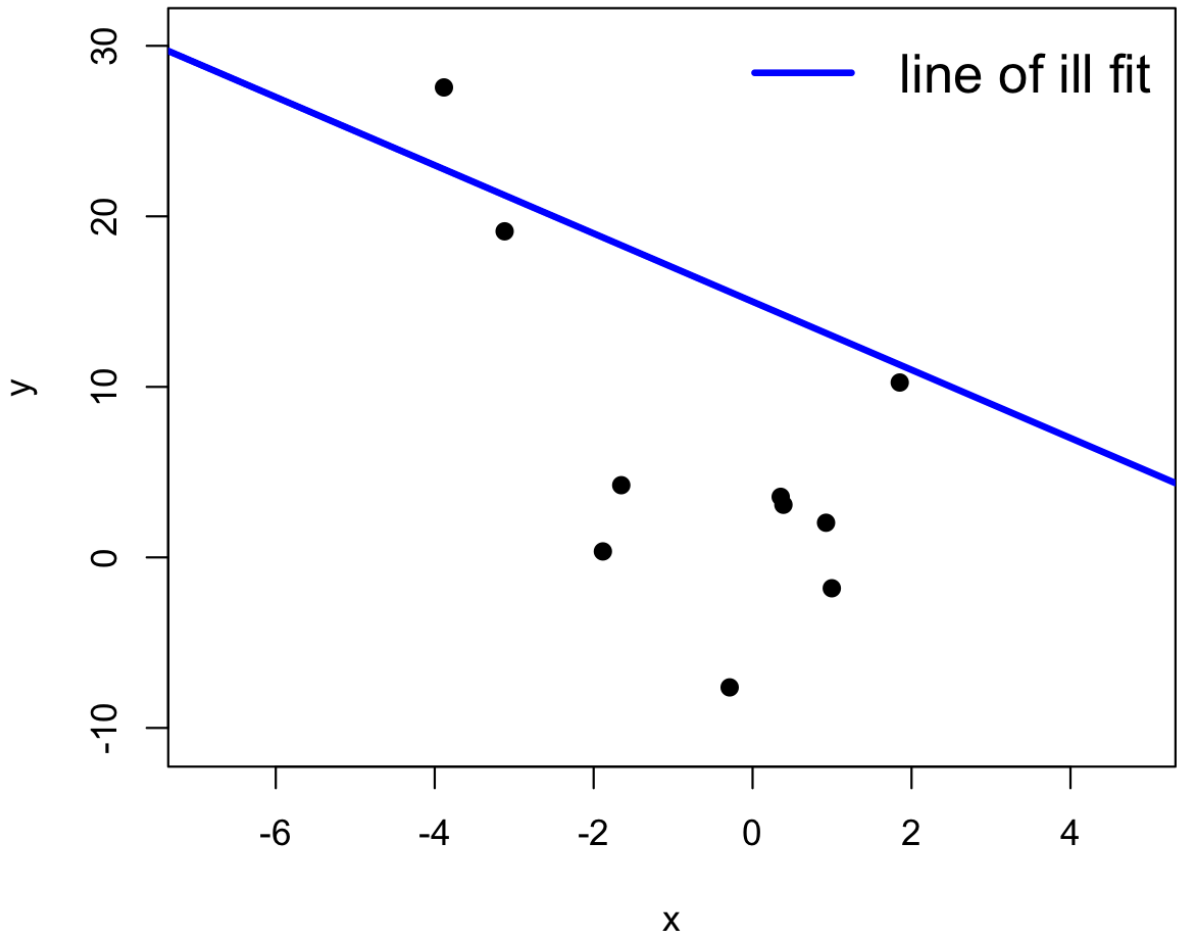
Turn messy, bivariate data into knowledge...



Question: how do we find this line? what do we mean by "best"?

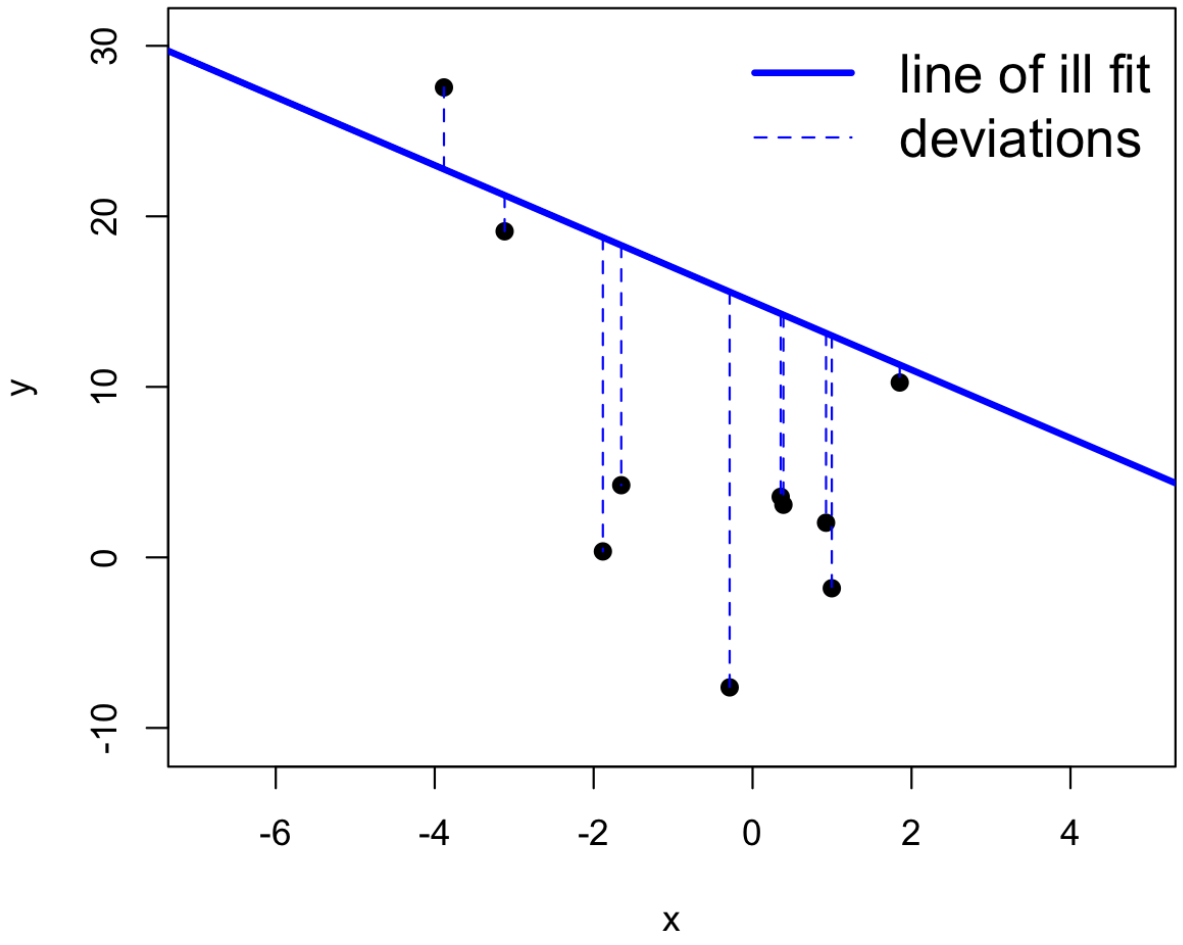
"Best" fit: minimizing deviations from the line

Any line we slap on the scatterplot will have errors:



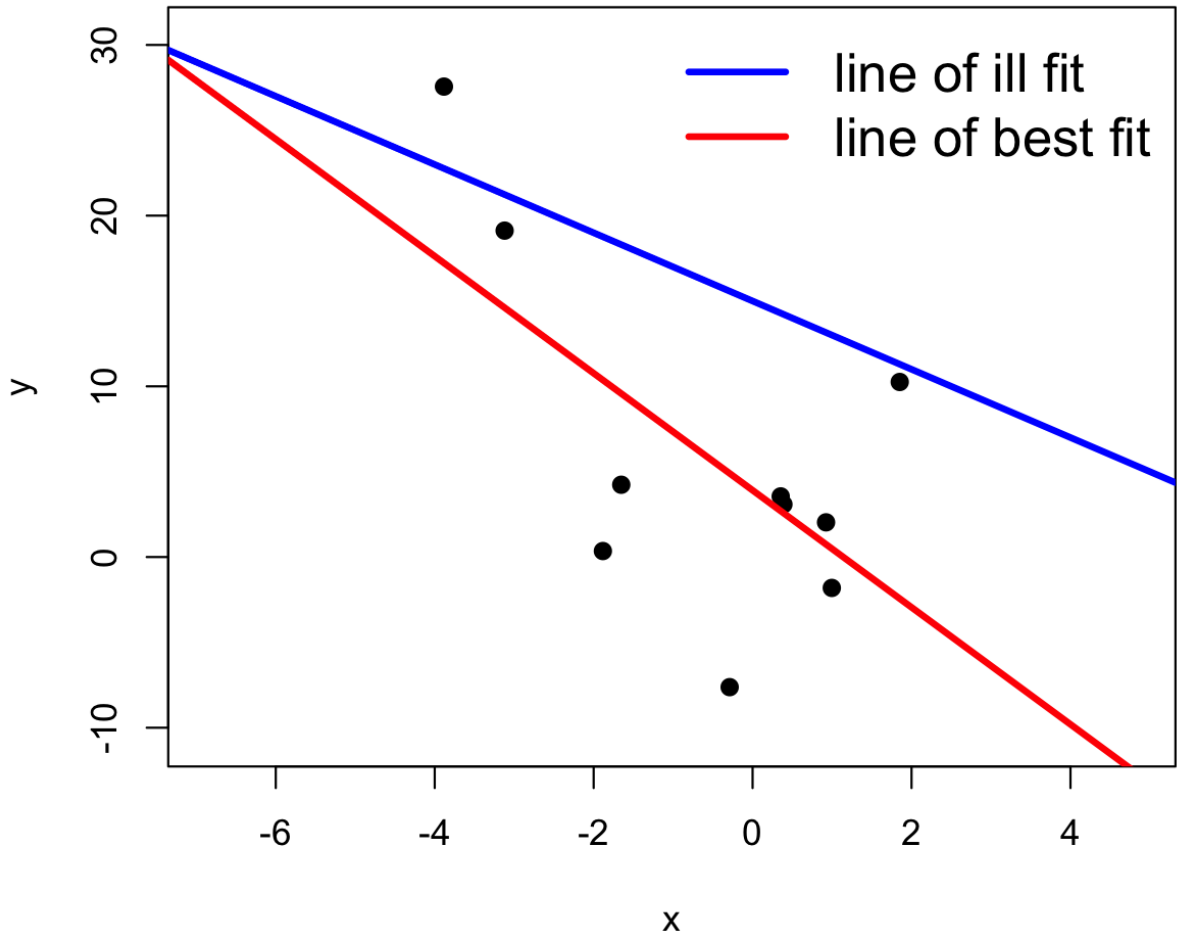
"Best" fit: minimizing deviations from the line

Any line we slap on the scatterplot will have errors:



"Best" fit: minimizing deviations from the line

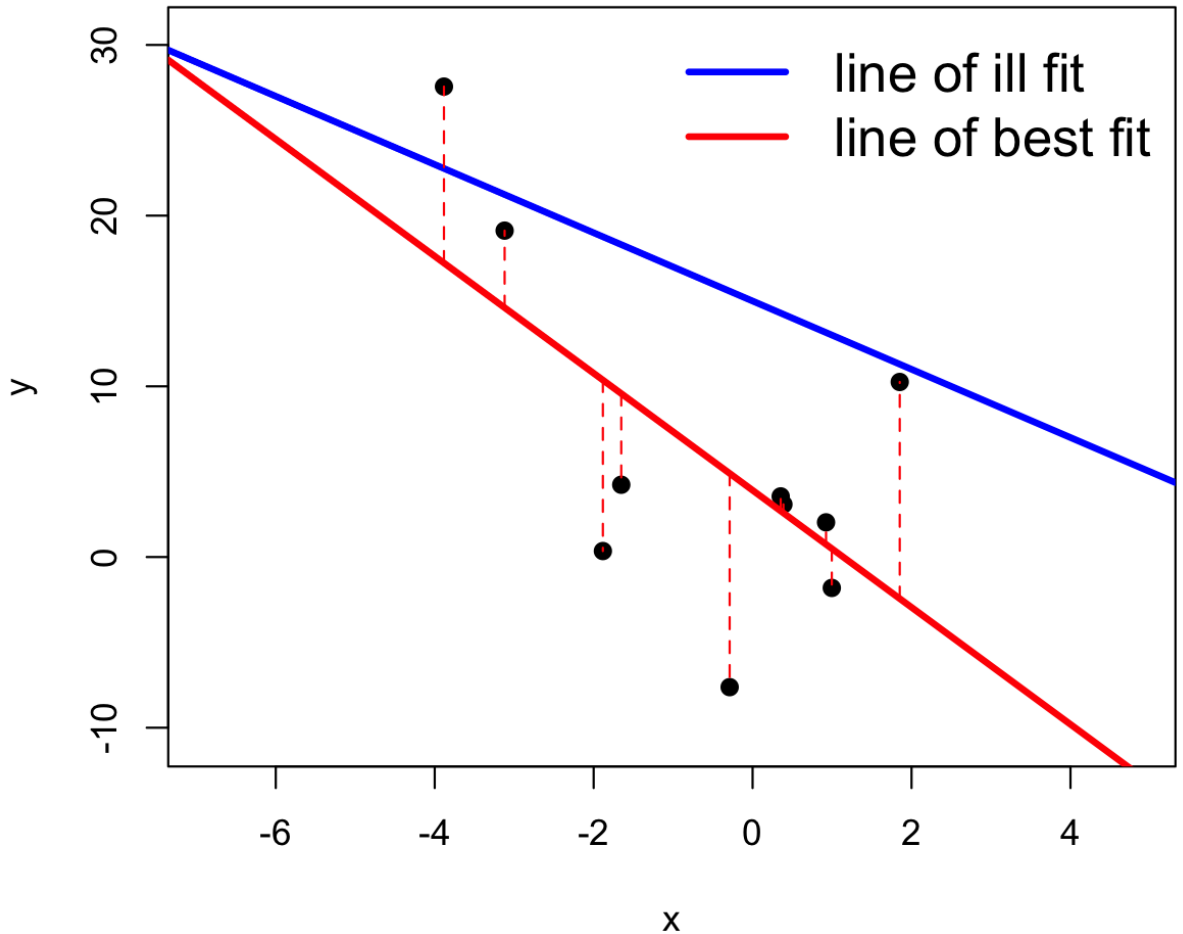
Any line we slap on the scatterplot will have errors:



The line of best fit is the line that makes the sum of these (squared) deviations as small as possible. Hence "best."

"Best" fit: minimizing deviations from the line

Any line we slap on the scatterplot will have errors:



The line of best fit is the line that makes the sum of these (squared) deviations as small as possible. Hence "best."

https://seeing-theory.brown.edu/regression-analysis

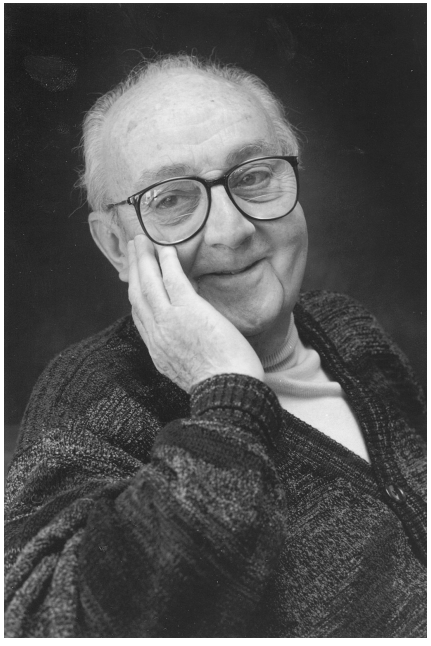
Our first statistical model: simple linear regression

$$y = \underbrace{\beta_0 + \beta_1 x}_{\text{fit/prediction}} + \underbrace{\varepsilon}_{\text{error}}$$

- y: *outcome* or *response* variable;
- x: *predictor*, *covariate*, *regressor*, *feature*, ...;
- β_0, β_1: *coefficients* or *parameters*;
- ε: *error* or *residual*;

This model predicts y given x.

Big ol' theme



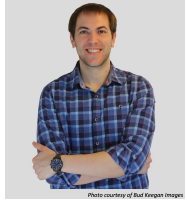
"All models are wrong but some are useful."

- George Box

↓


→

→

→

→


Sample vs population

Sample: finite, incomplete, real-world dataset you actually have;

Population: infinite, complete, idea dataset you wish you had;

The "ideal" model:

$$y = \beta_0 + \beta_1 x + \varepsilon.$$

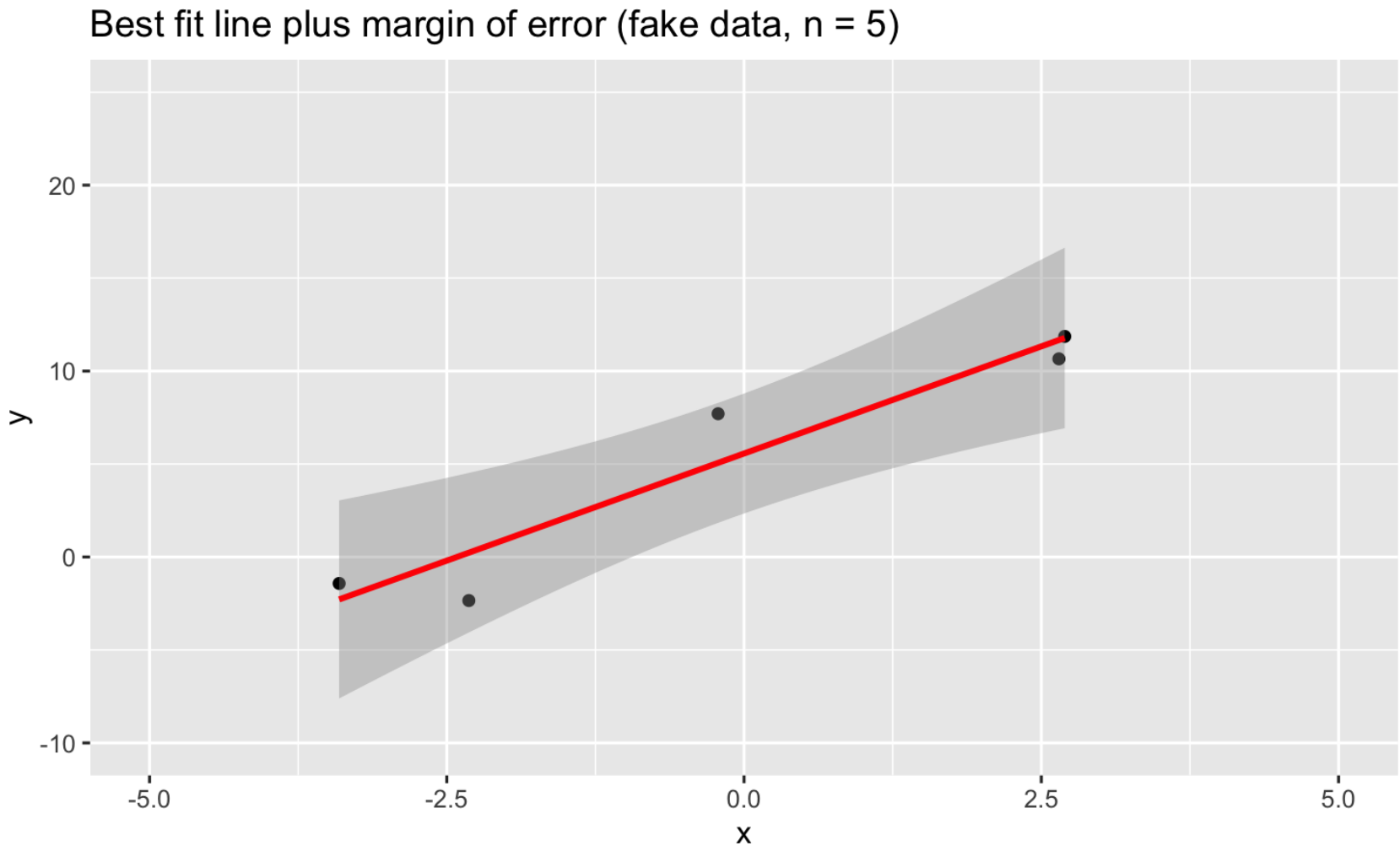
The fitted or estimated line is:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x.$$

$\hat{\beta}_0$, $\hat{\beta}_1$ are our *estimate* ("best guess") based on the imperfect sample, and we hope they are close to the ideal values β_0, β_1, but there will be uncertainty (margin of error).

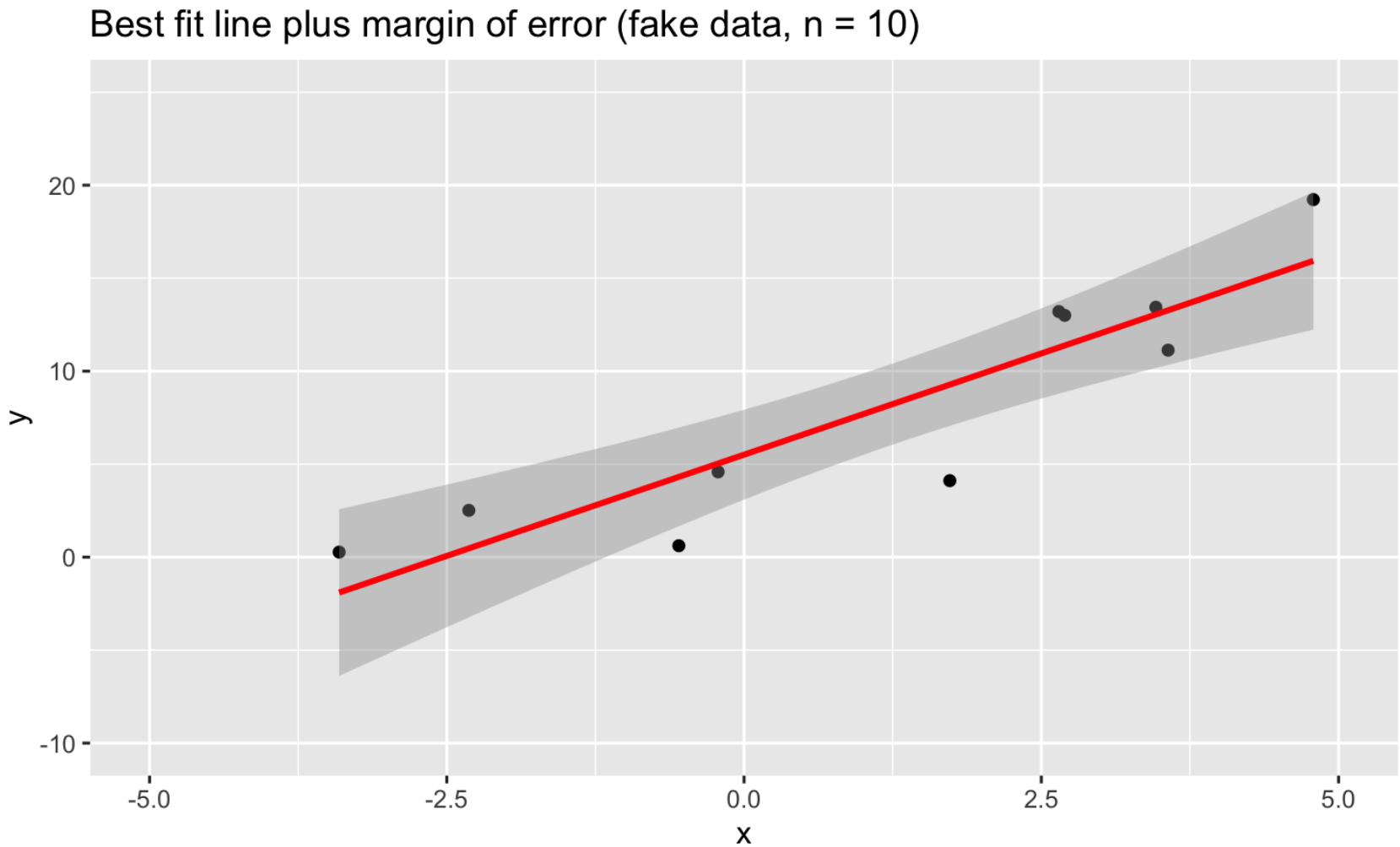
We calculate $\hat{\beta}_0$, $\hat{\beta}_1$ by minimizing squared errors like you saw before. They are "least squares" estimates.

Estimates getting better and better



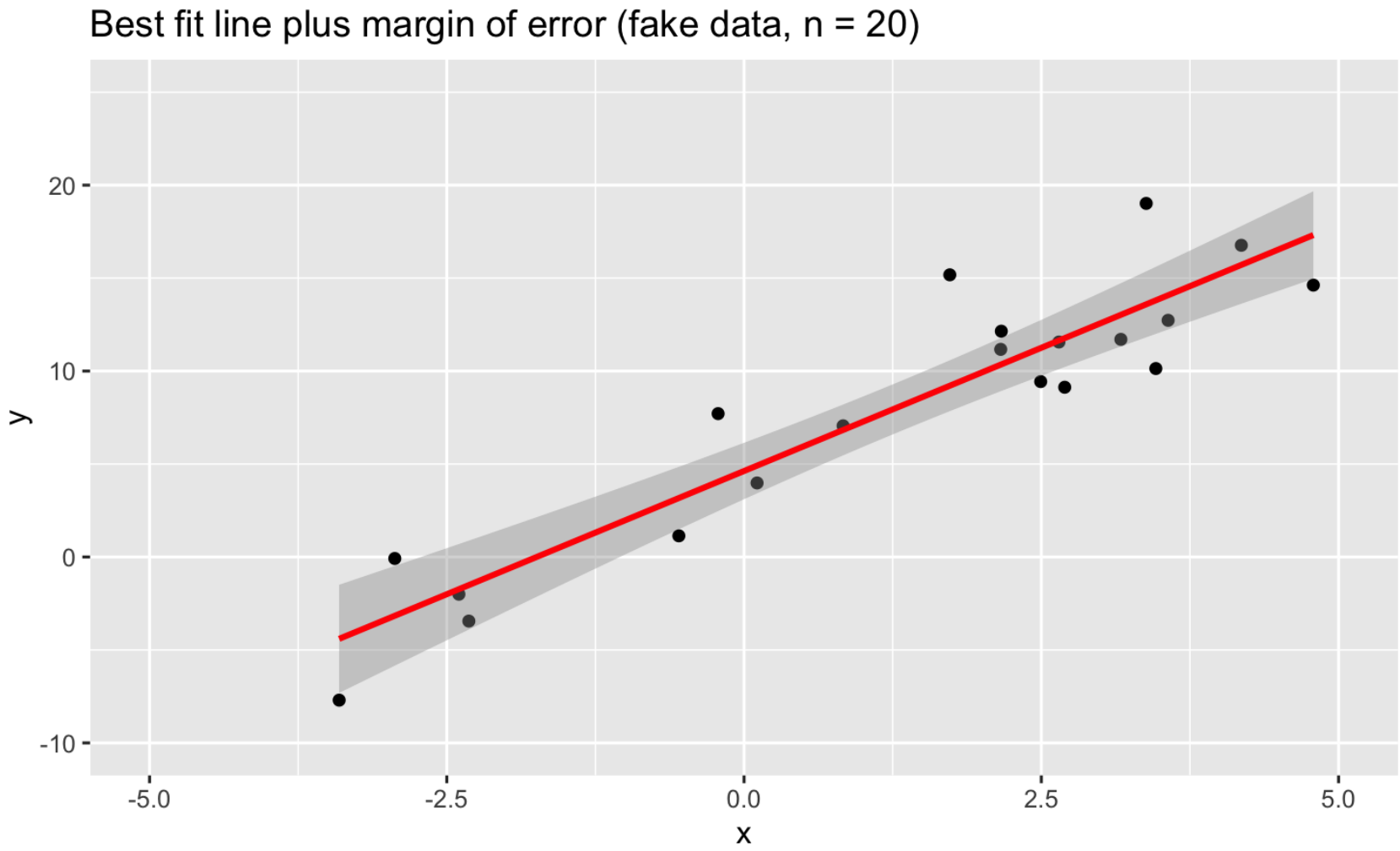
This is an idealized "laboratory" where the only thing wrong with our data is that we don't have enough of it.

Estimates getting better and better



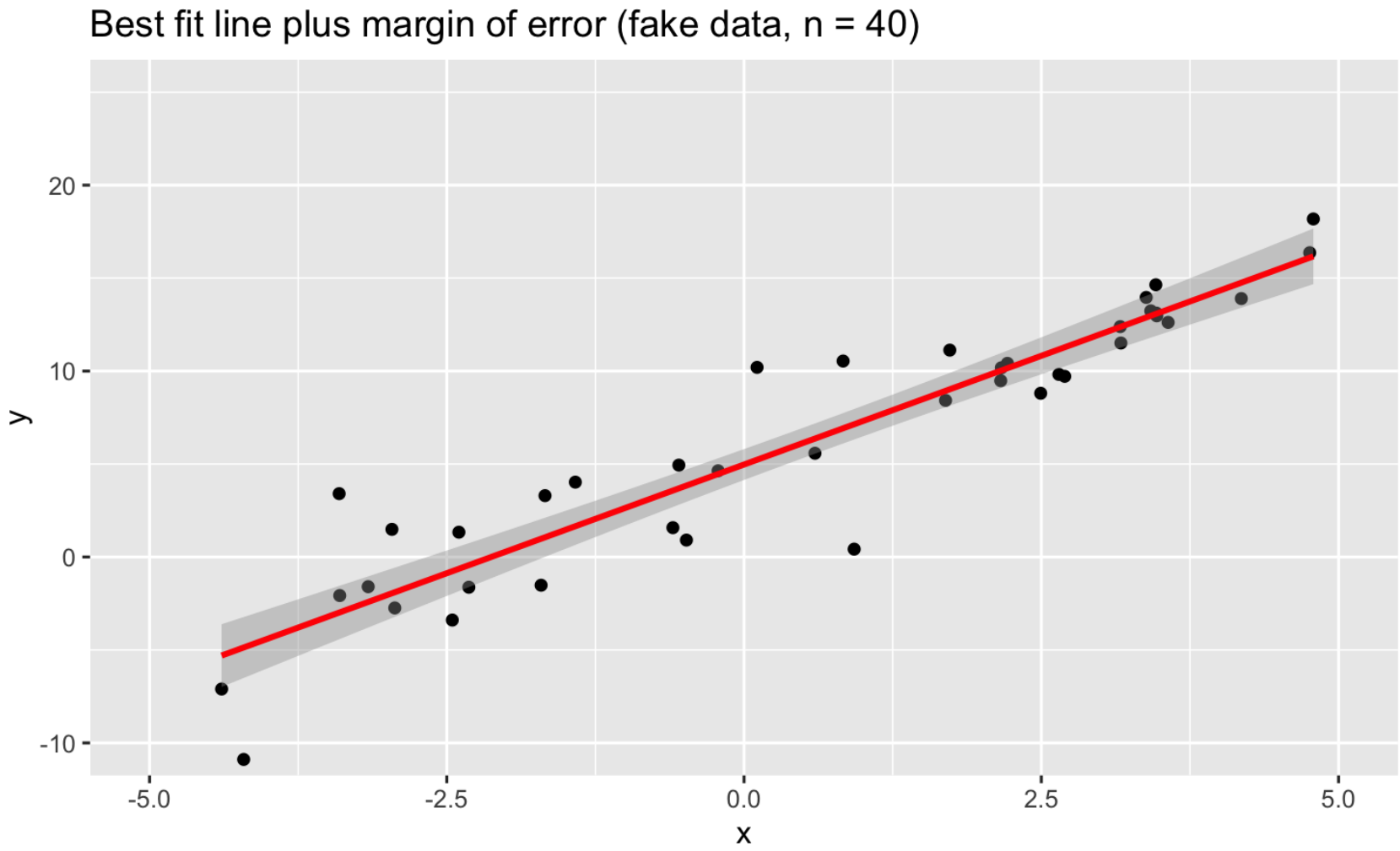
This is an idealized "laboratory" where the only thing wrong with our data is that we don't have enough of it.

Estimates getting better and better



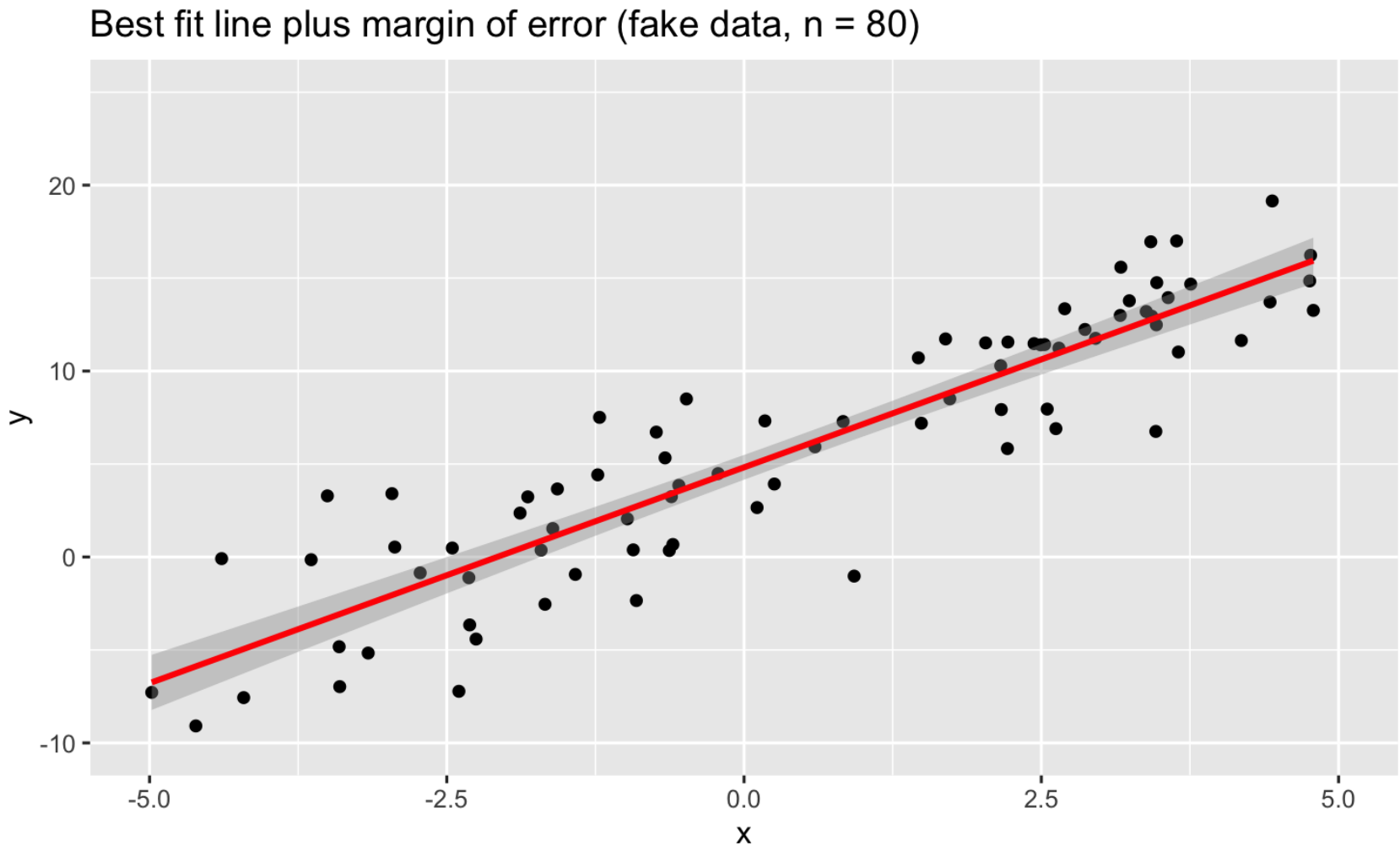
This is an idealized "laboratory" where the only thing wrong with our data is that we don't have enough of it.

Estimates getting better and better



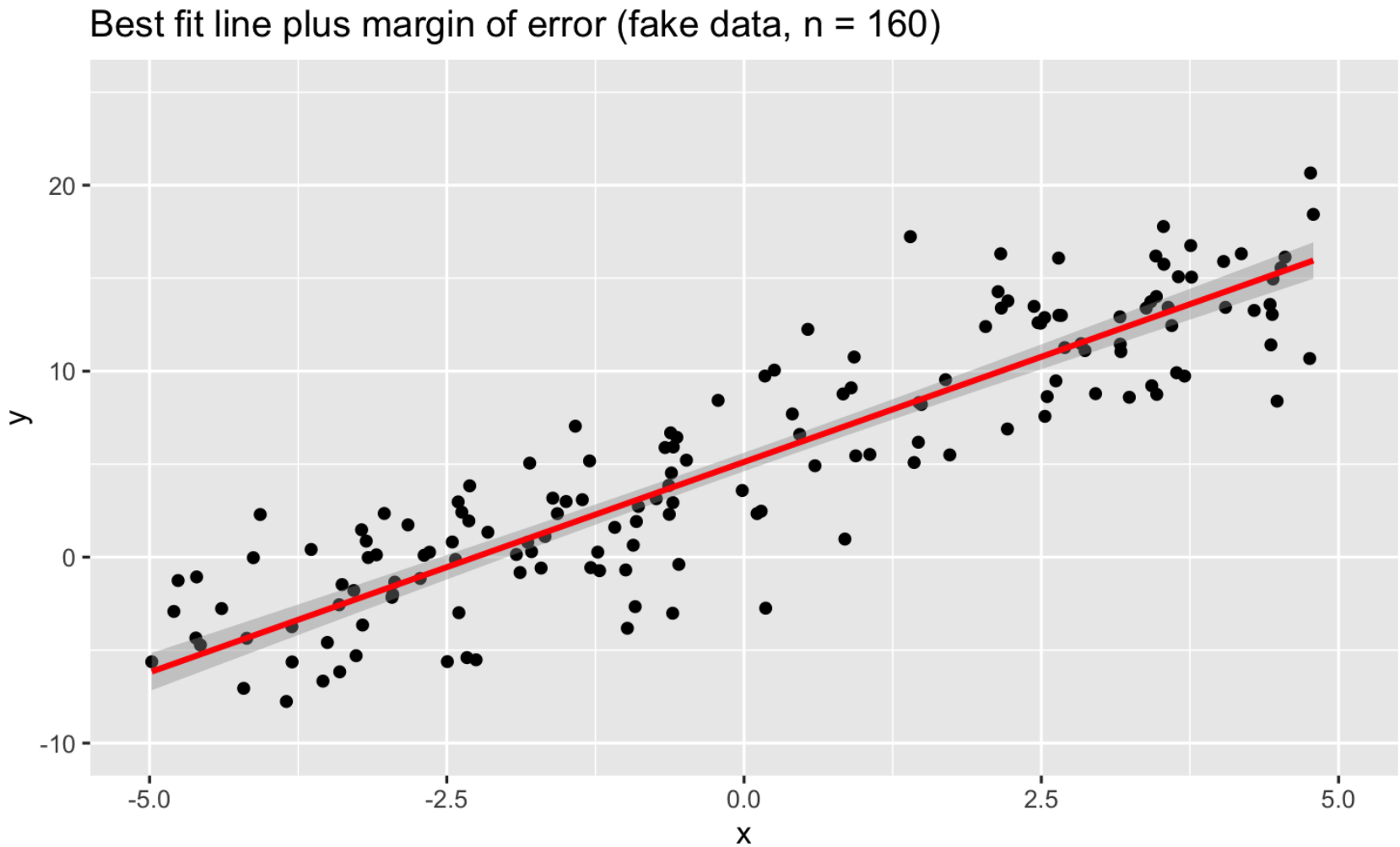
This is an idealized "laboratory" where the only thing wrong with our data is that we don't have enough of it.

Estimates getting better and better



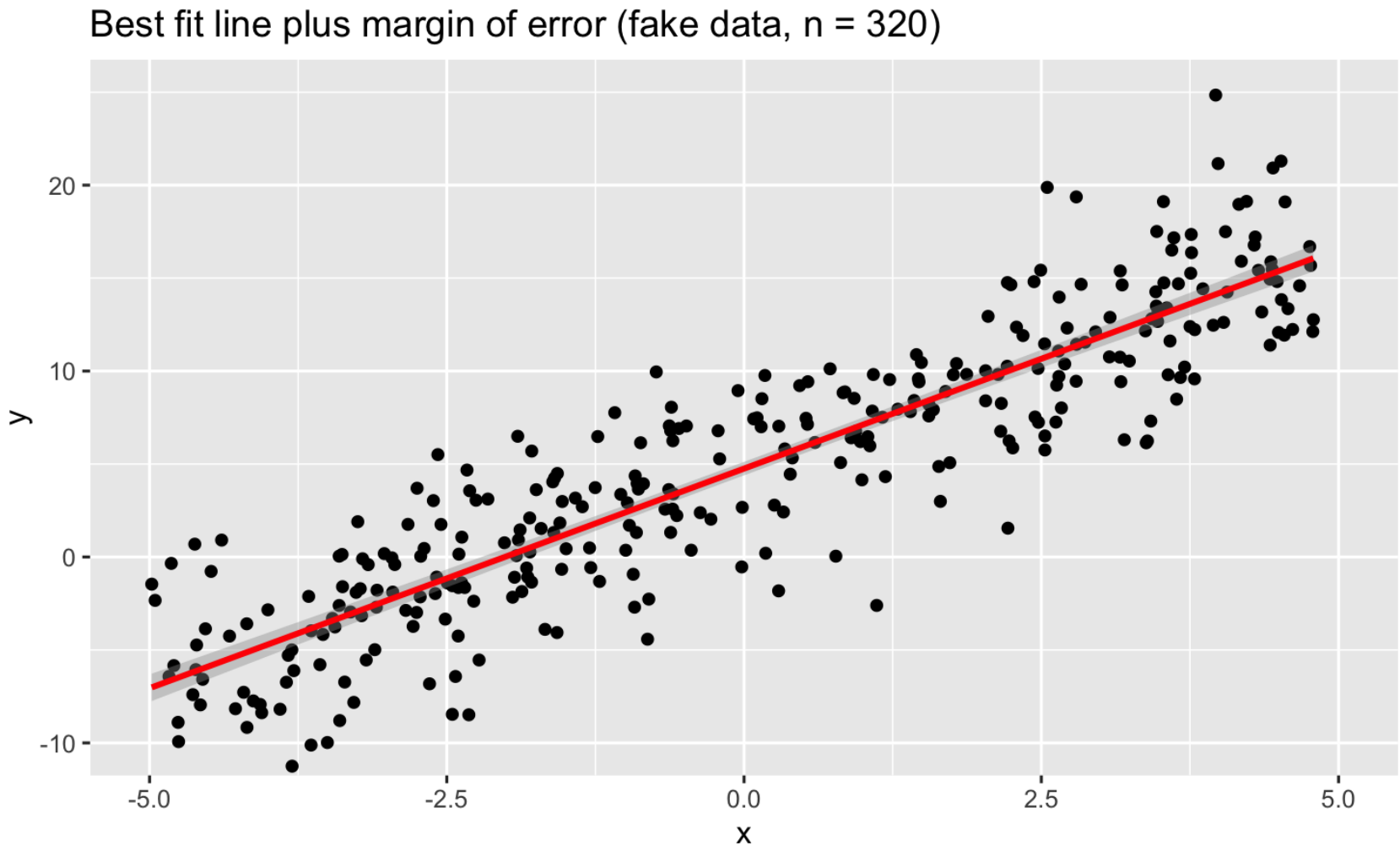
This is an idealized "laboratory" where the only thing wrong with our data is that we don't have enough of it.

Estimates getting better and better



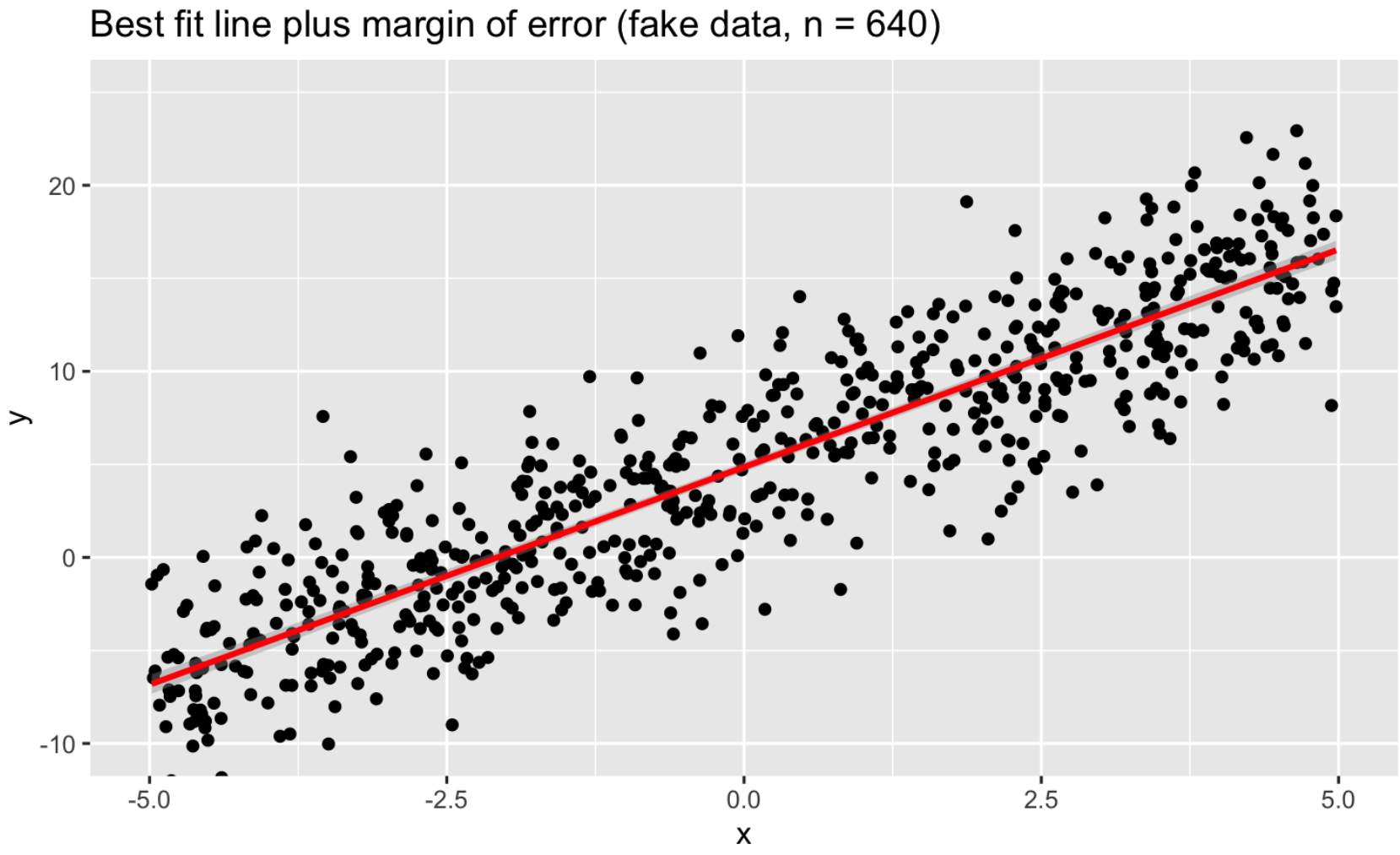
This is an idealized "laboratory" where the only thing wrong with our data is that we don't have enough of it.

Estimates getting better and better



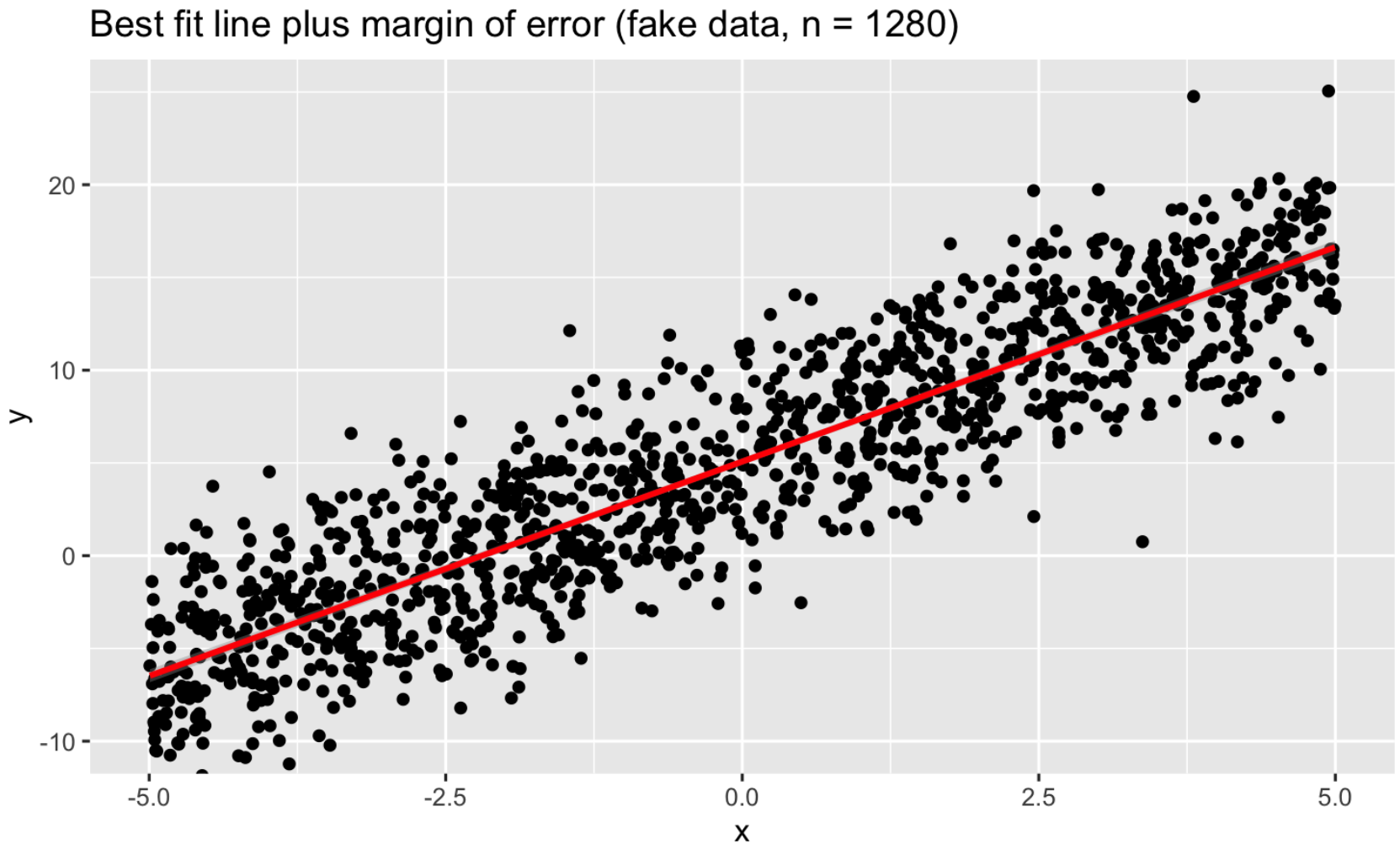
This is an idealized "laboratory" where the only thing wrong with our data is that we don't have enough of it.

Estimates getting better and better



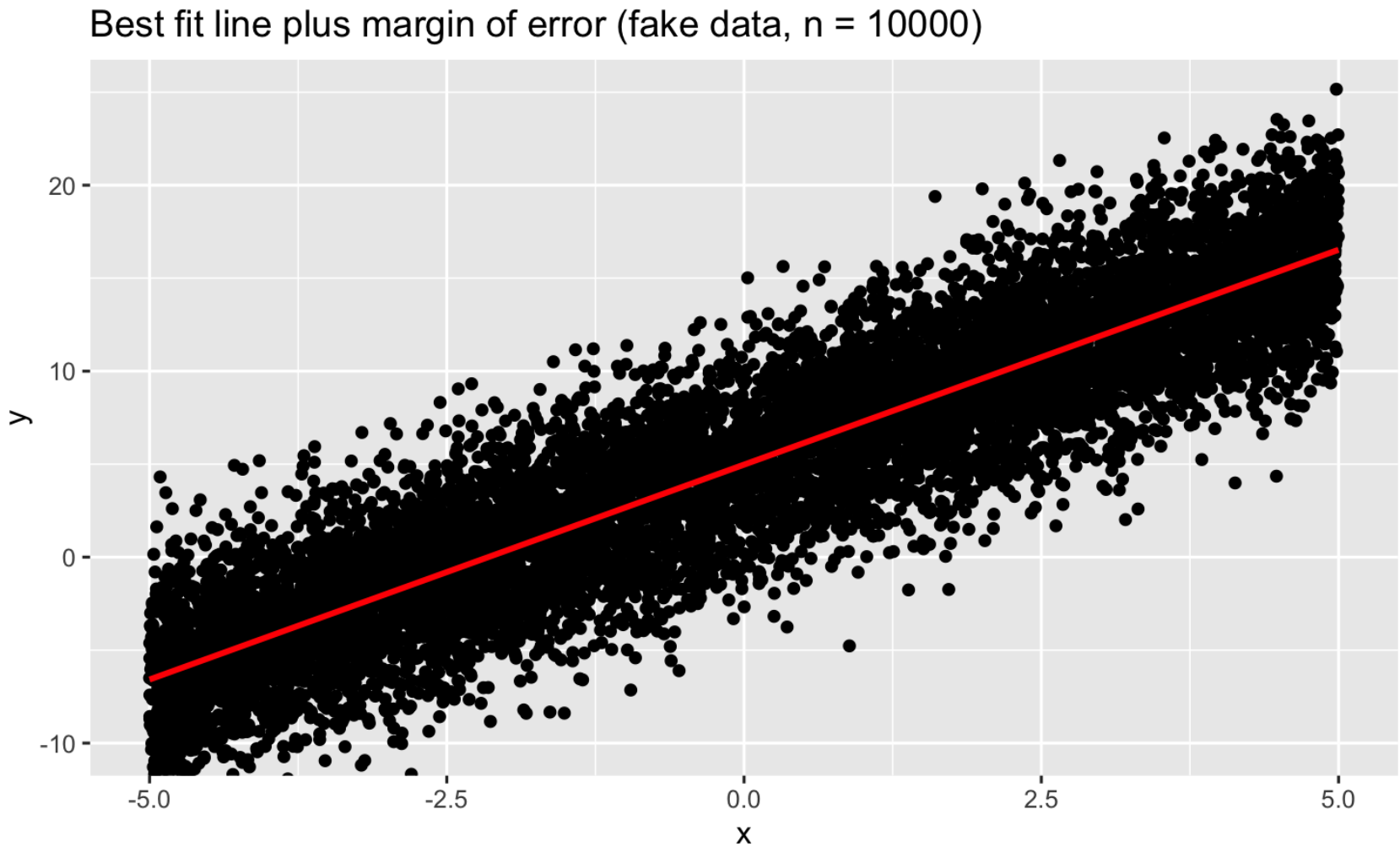
This is an idealized "laboratory" where the only thing wrong with our data is that we don't have enough of it.

Estimates getting better and better



This is an idealized "laboratory" where the only thing wrong with our data is that we don't have enough of it.

Estimates getting better and better



This is an idealized "laboratory" where the only thing wrong with our data is that we don't have enough of it.

Predictions

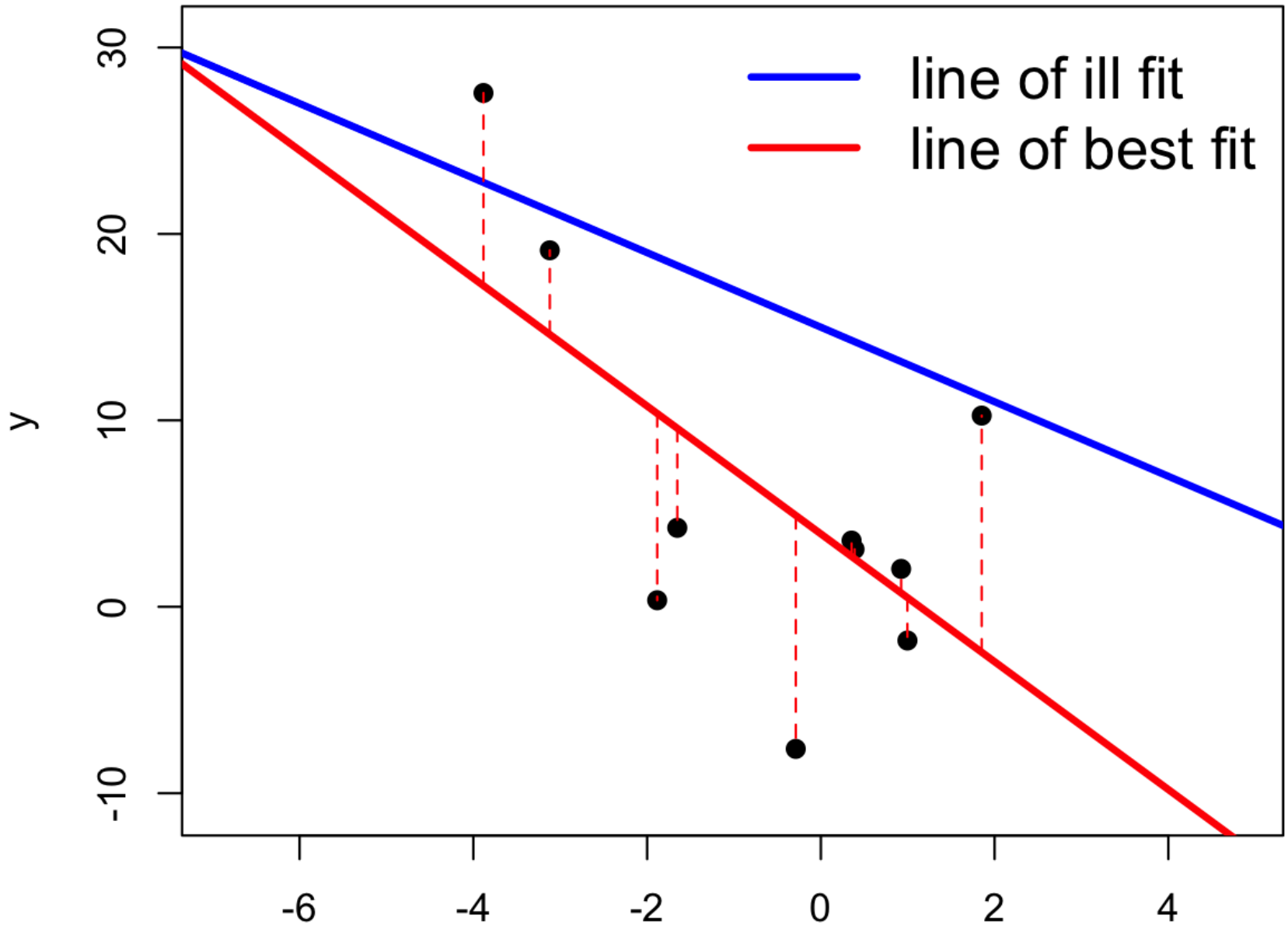
- If you give me an x, I can use the model to predict what I think y will be:
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x;$$

- If we do this for an x value that is already in my dataset, then $\hat{y}$ is called the *fitted value*;

- The difference between what the model predicts and what actually happened in the data is the residual:
$$\hat{\varepsilon} = y - \hat{y} = y - \hat{\beta}_0 - \hat{\beta}_1 x.$$

  The vertical deviation between the data and the line!

The vertical deviations are $\hat{\varepsilon}$



Estimation

data			residuals
x_1	y_1	$\rightarrow$	$\hat{\varepsilon}_1 = y_1 - \hat{y}_1$
x_2	y_2	$\rightarrow$	$\hat{\varepsilon}_2 = y_2 - \hat{y}_2$
x_3	y_3	$\rightarrow$	$\hat{\varepsilon}_3 = y_3 - \hat{y}_3$
x_4	y_4	$\rightarrow$	$\hat{\varepsilon}_4 = y_4 - \hat{y}_4$

$$\downarrow$$

$$\underbrace{\hat{\varepsilon}_1^2 + \hat{\varepsilon}_2^2 + \hat{\varepsilon}_3^2 + \hat{\varepsilon}_4^2}_{\text{sum of squared residuals}}$$

We pick $\hat{\beta}_0$, $\hat{\beta}_1$ so that $\sum \hat{\varepsilon}_i^2$ is as small as possible ("best fit").