Bootstrap Confidence Intervals

ST552 Lecture 13

Charlotte Wickham

2019-02-11

Motivation

The inferences we've covered so far relied on our assumption of Normal errors:

$$\epsilon \sim N(0, \sigma^2 I_{n \times n})$$

For example, we've seen under this assumption, the least squares estimates are also Normally distributed:

$$\hat{\beta} \sim N\left(\beta, \sigma^2 \left(X^T X\right)^{-1}\right)$$

If the errors aren't truly Normally distributed, what distribution do the estimates have?

Warm-up: Your Turn

Imagine the errors are in fact t_3 distributed?

With your neighbour: **design a simulation to understand the distribution of the least squares estimates.**

sampling

a particular β

Some model $\varepsilon \sim t_3$

① Assume some things about model: $y = X\beta + \varepsilon$

Decide on n, p, X and β, make something up.

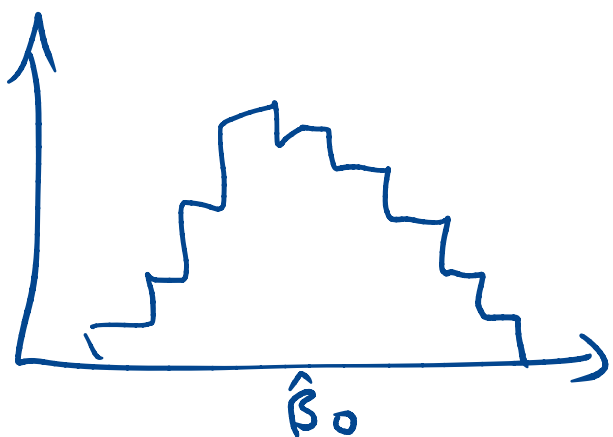
② Compute $\hat{\beta}$'s 100,000 times

a) Simulate y: simulate ε freely, $y = X\beta + \varepsilon$

b) Using y, fit model: $\hat{\beta} = (X^T X)^{-1} X^T y$

③ Make Histogram

Examine many $\hat{\beta}$'s

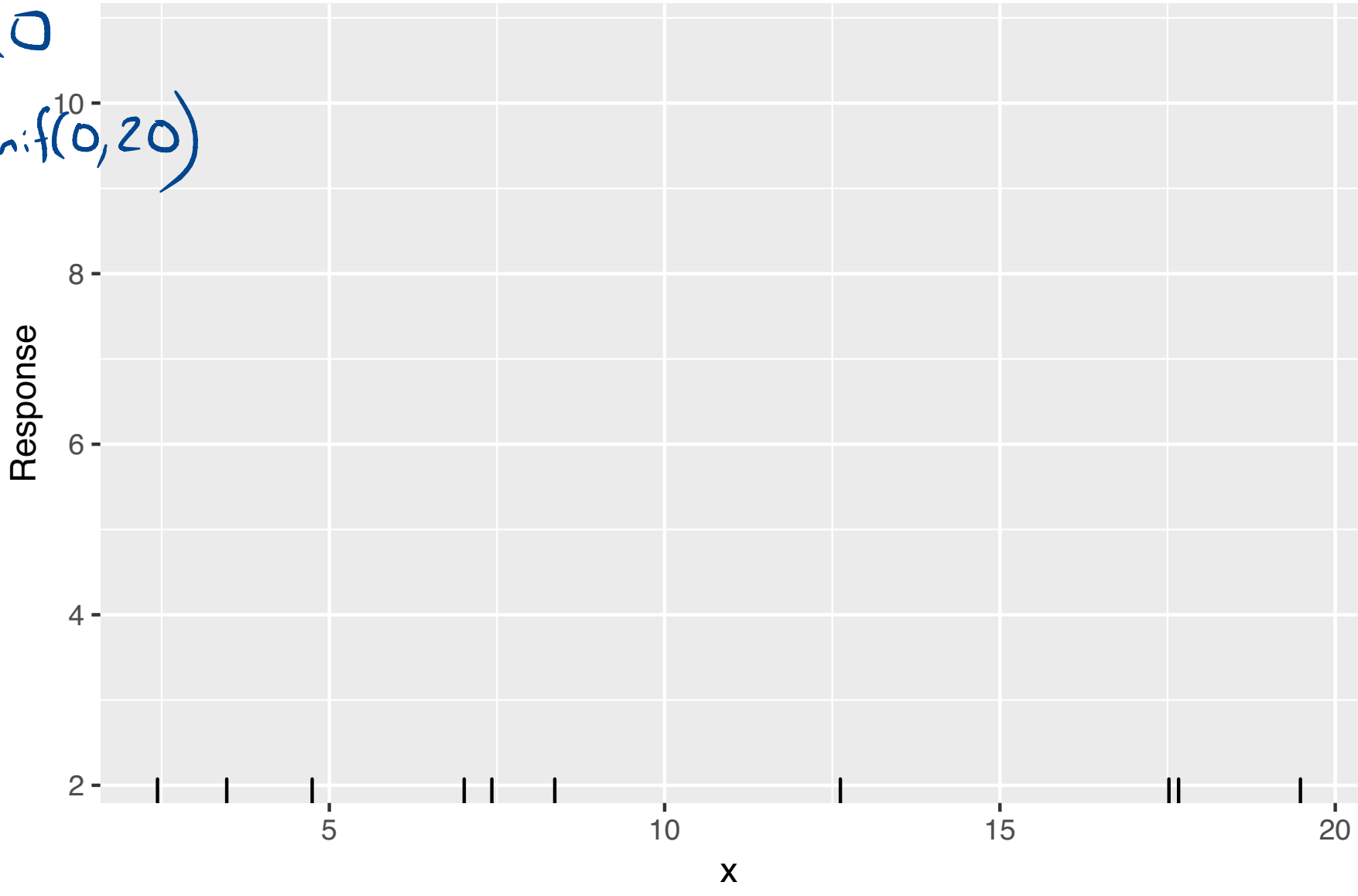


Rod: We don't know this with real data

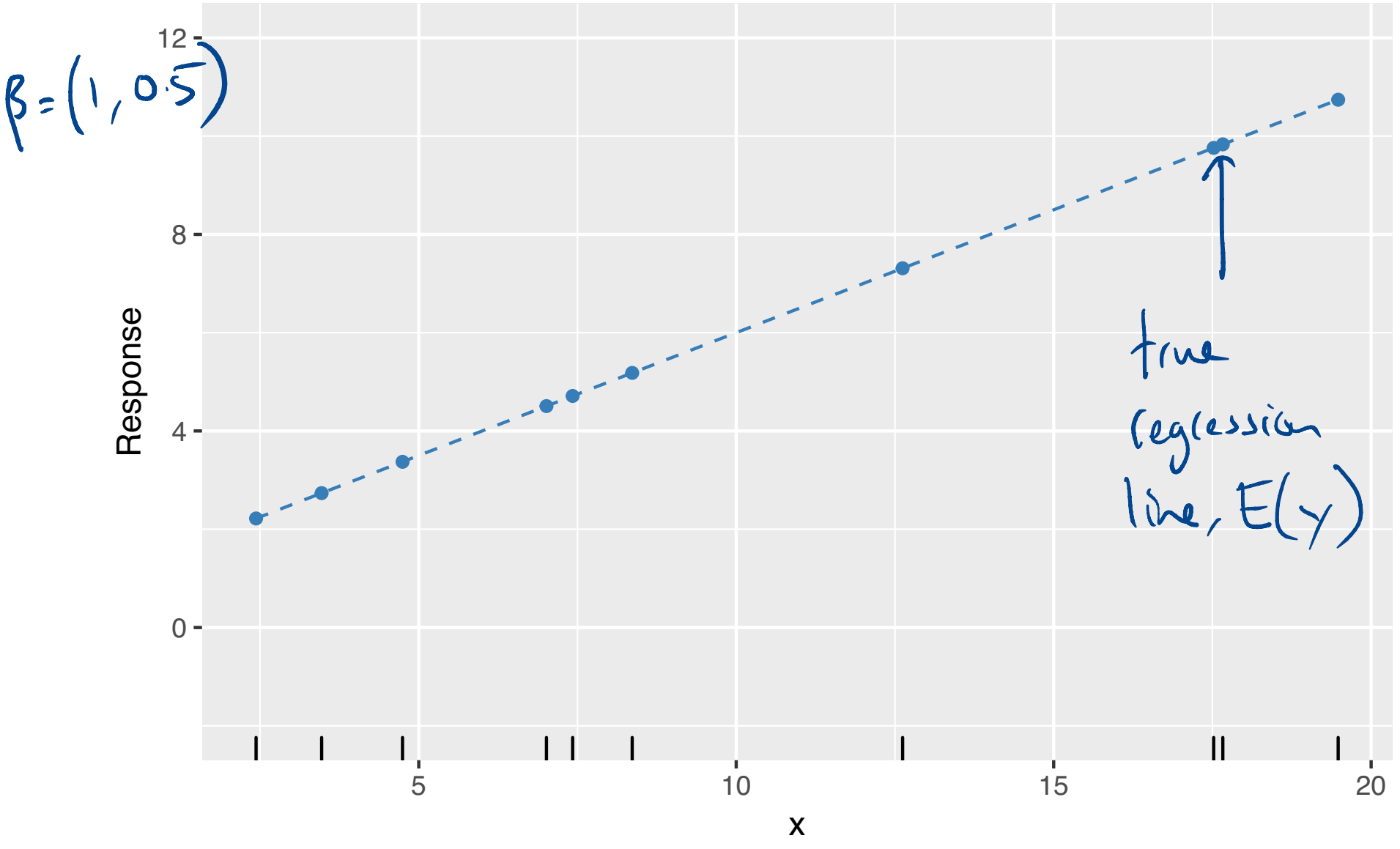
Example: 1. Fix n, fix X

$n = 10$

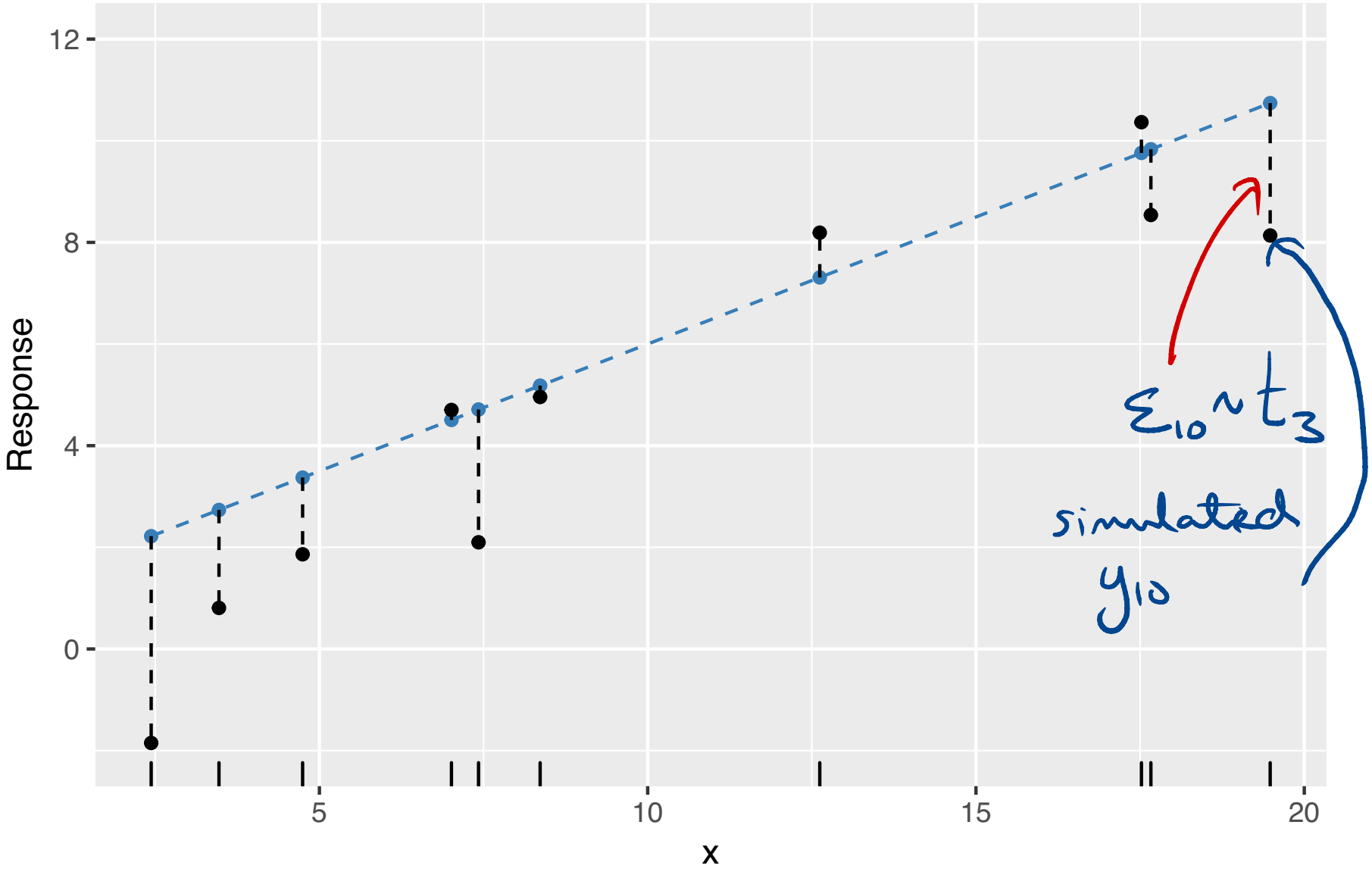
$X \sim \text{Unif}(0, 20)$



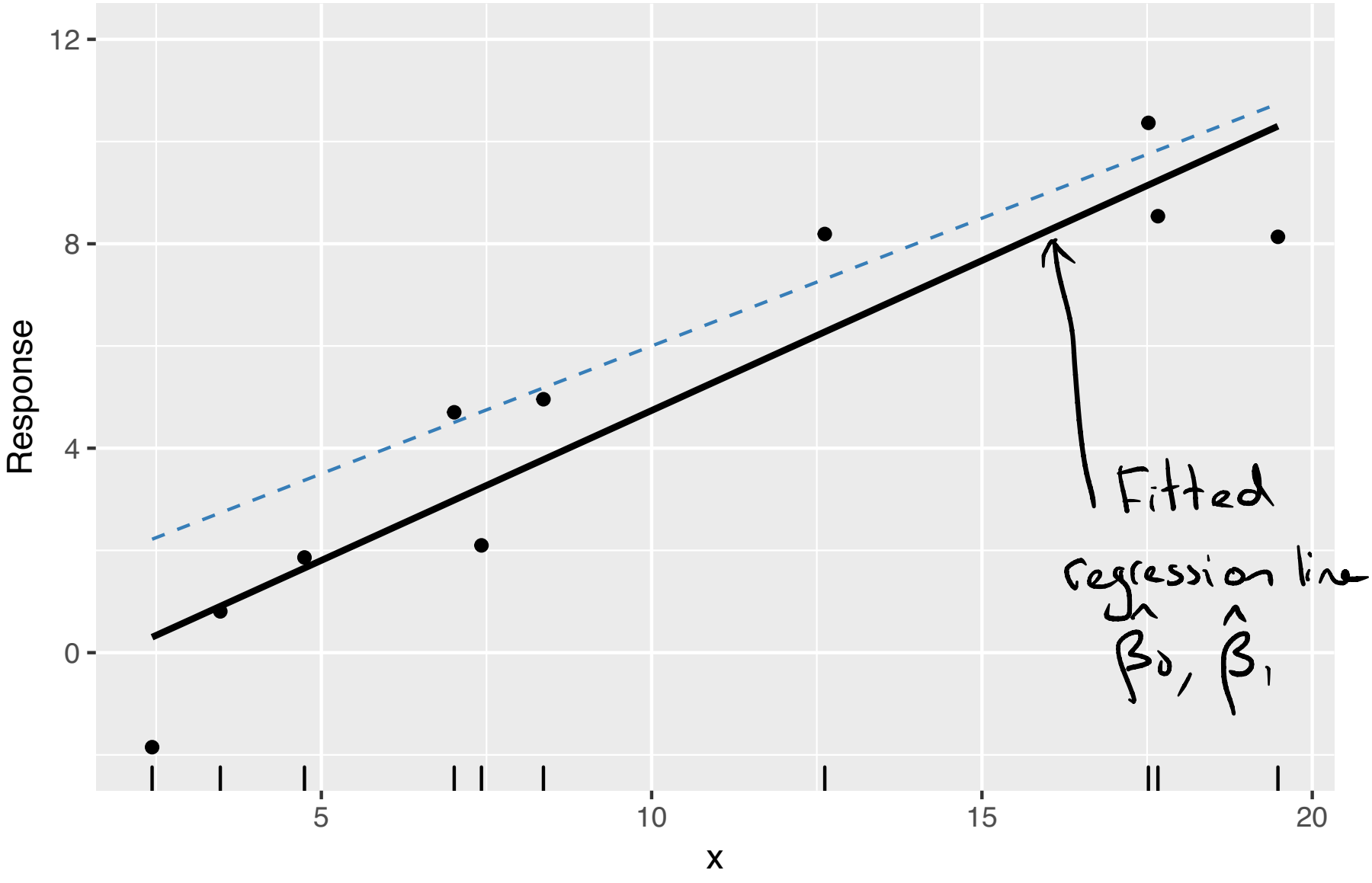
Example: 1. Fix β, find $\hat{y}$ ~~X~~ $E(y)$



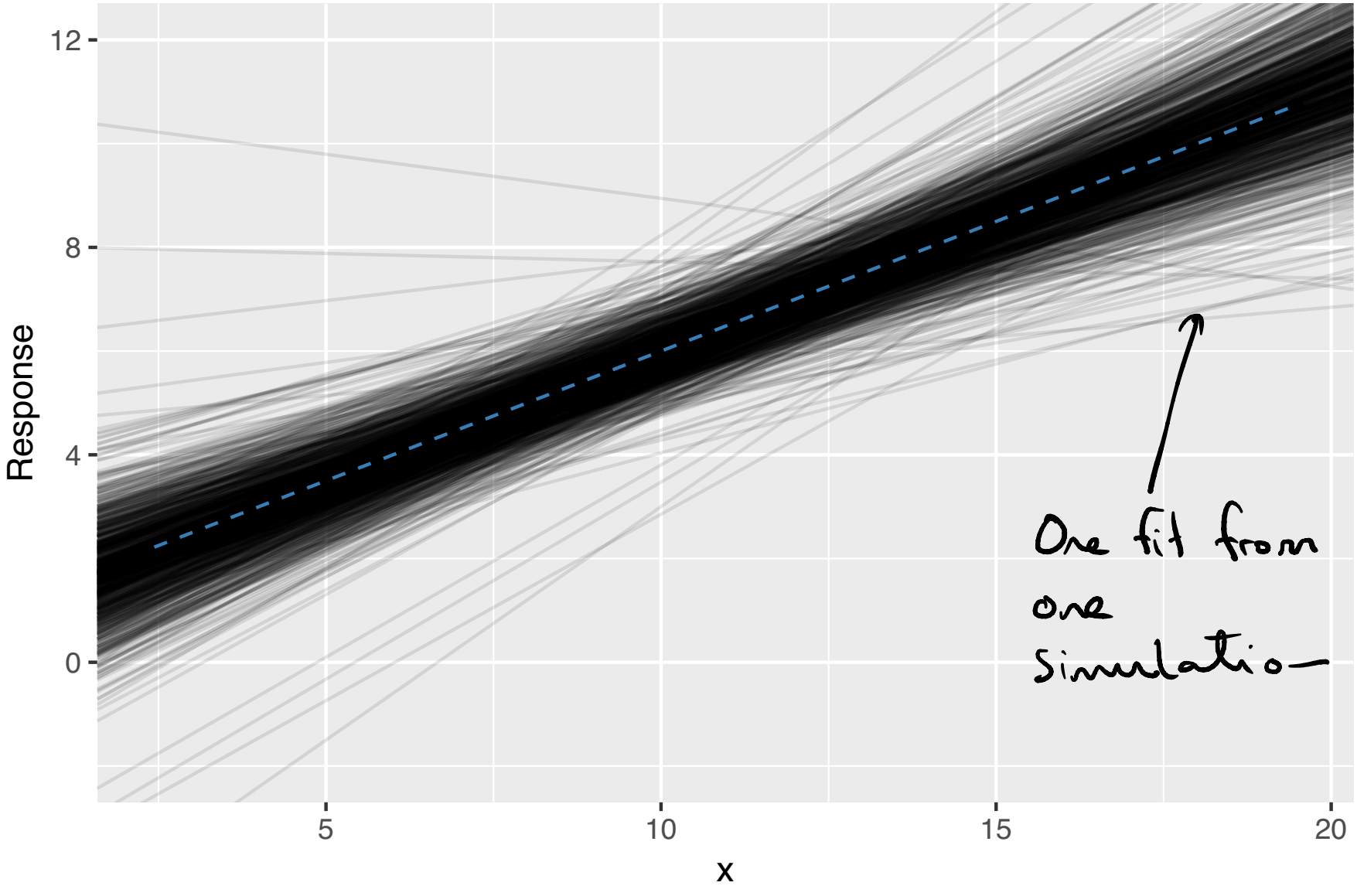
Example: 2. Simulate errors, find y



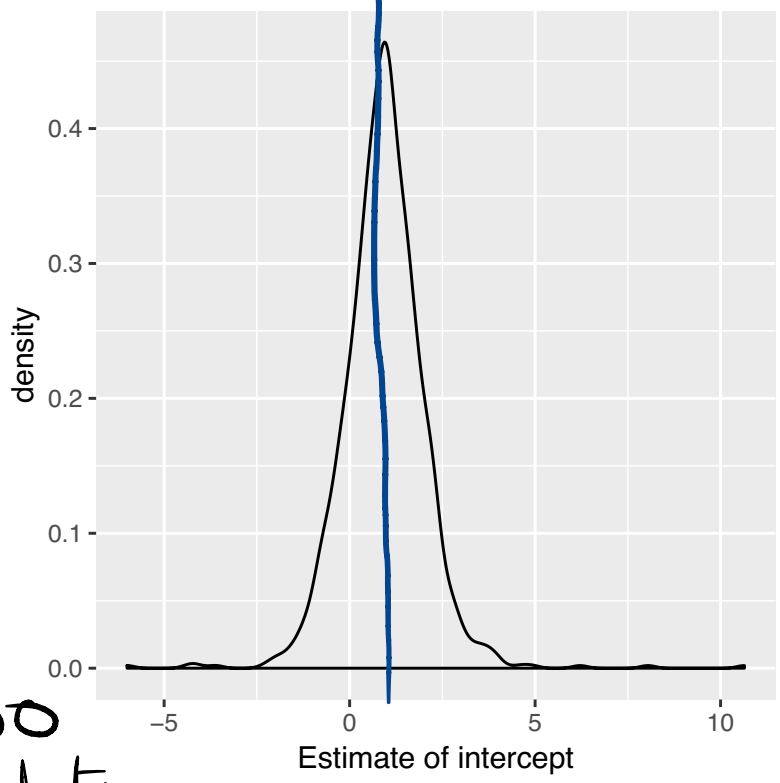
Example: 3. Find least squares line



Example: 4. Repeat #2. and #3. many times



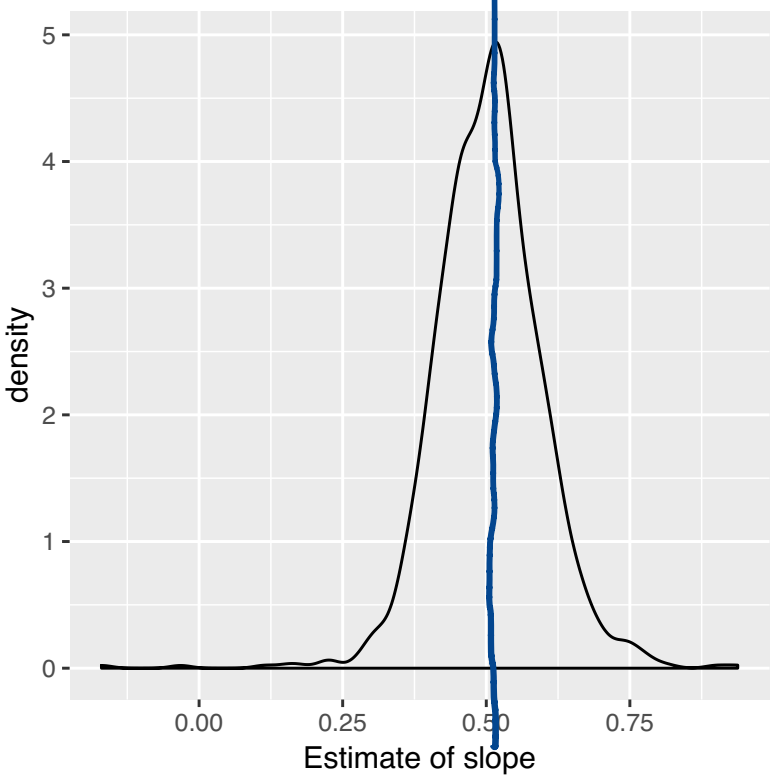
Example: Examine distribution of estimates



$\beta_0 = 1$

1000 simulations

1000 $\hat{\beta}_0$

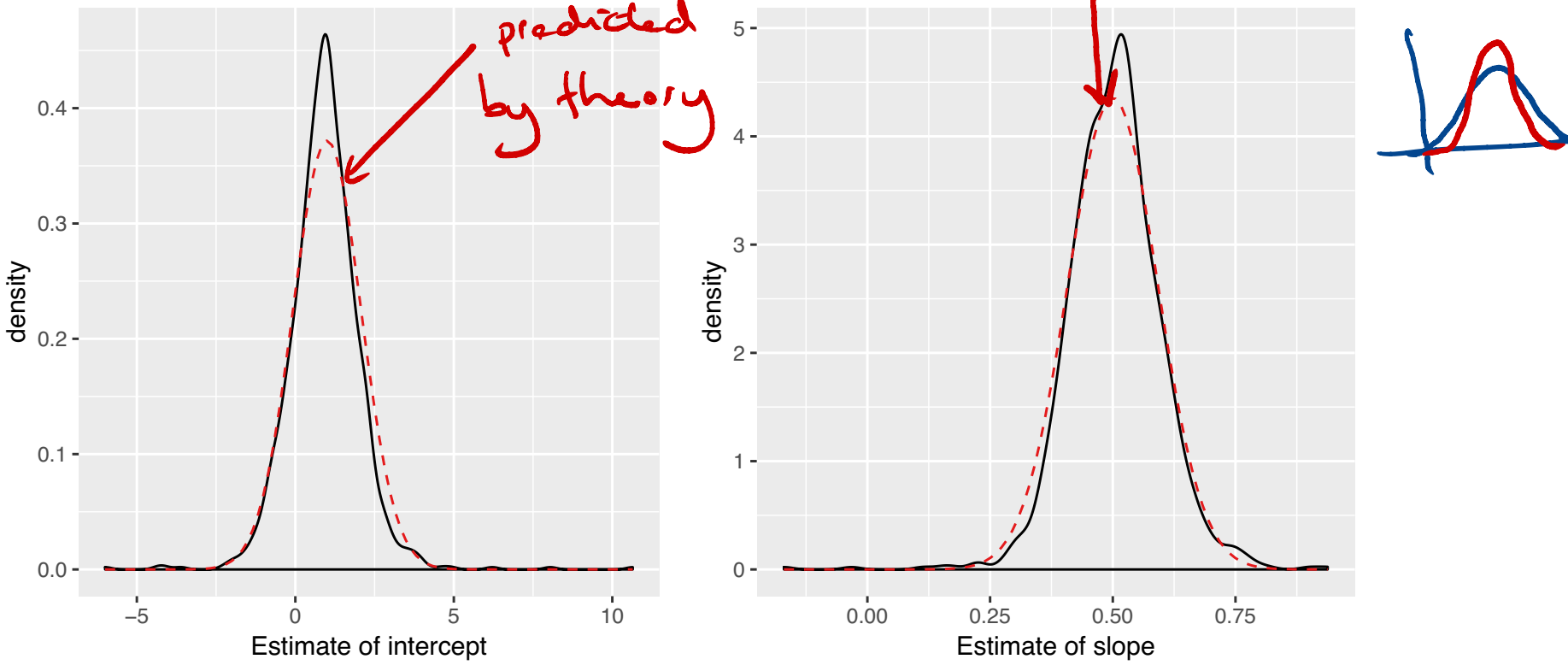


$\beta_1 = 0.5$

1000 $\hat{\beta}_1$

Example: Compared to theory

If $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$ $\quad \hat{\beta} \sim N(\beta, \sigma^2 (X^T X)^{-1})$



Based on simulation:

If error are t_3 distributed looks close to Normal.

We assumed on n, X, β.

When the errors aren't Normal: CLT

$\hat{\beta} = \_\_\_ + \_\_\_ \varepsilon$

Think of our estimates like linear combinations of the errors. I.e. a sort of average of i.i.d random variables.
Some version of the Central Limit Theorem will apply.

For large samples, even when the errors aren't Normal,

$$\hat{\beta} \dot{\sim} N(\beta, \sigma^2 (X^T X)^{-1})$$

approximately

as $n \to \infty$ that approximation will improve.

Summary so far

If we knew the error distribution and true parameters we could use simulation to understand the sampling distribution the least squares estimates.

Simulation can also be used to demonstrate the CLT at work in regression.

Bootstrap confidence intervals

In practice, with data in front of us, we don't know the distribution of the errors (nor the true parameter values).

The bootstrap is one approach to estimate the sampling distribution of $\hat{\beta}$, by using the simulation idea, and substituting in our best guesses for the things we don't know.

Bootstrapping regression

(Model based resampling)

0. Fit model and find estimates, $\hat{\beta}$, and residuals, e_i

1. Fix X,
2. For $k = 1, \ldots, B$ Repeat many times
   - 2.1 Generate errors, ϵ_i^* sampled with replacement from e_i (residuals)
   - 2.2 Construct y, using the model, $y = \hat{y} + \epsilon^* = X\hat{\beta} + \epsilon^*$
   - 2.3 Use least squares to find $\hat{\beta}^*_{(k)}$
3. Examine the distribution of $\hat{\beta}^*$ and compare to $\hat{\beta}$

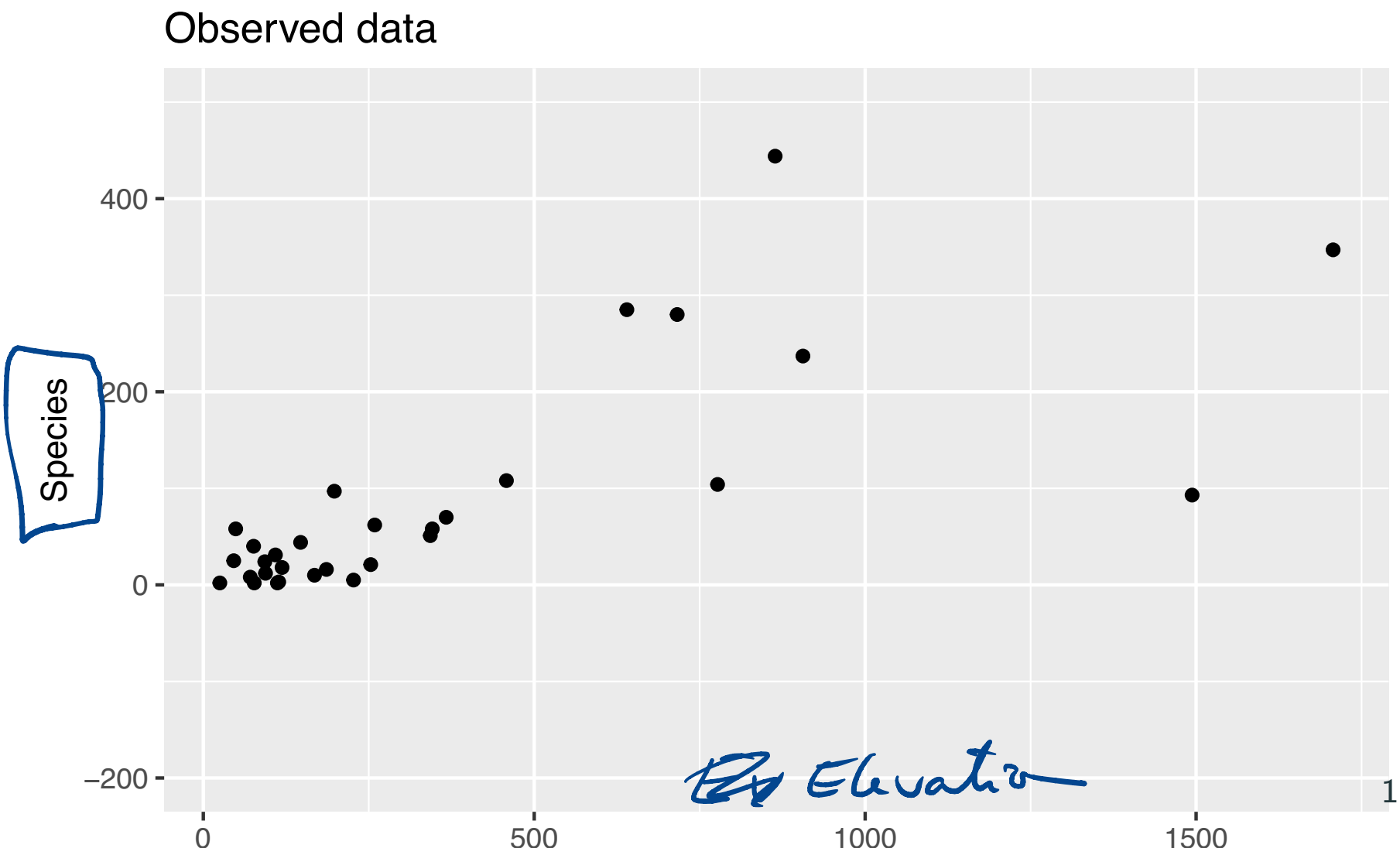
95%.

One confidence interval for β_j is the 2.5% and 97.5% quantiles of the distribution of $\hat{\beta}^*_j$.

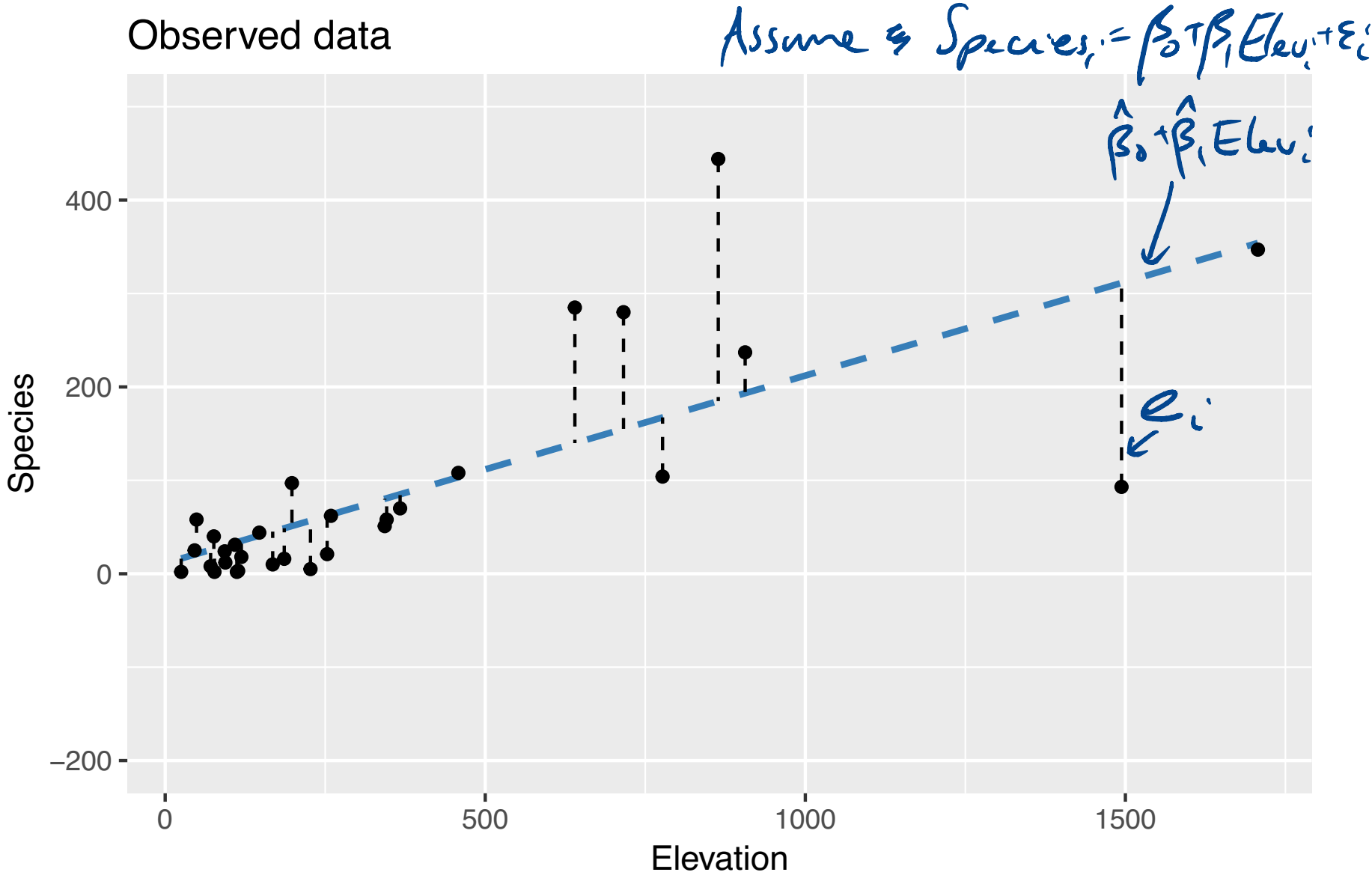
(Known as the Percentile method, there are other (better?) methods).

Example: Faraway Galapagos Islands

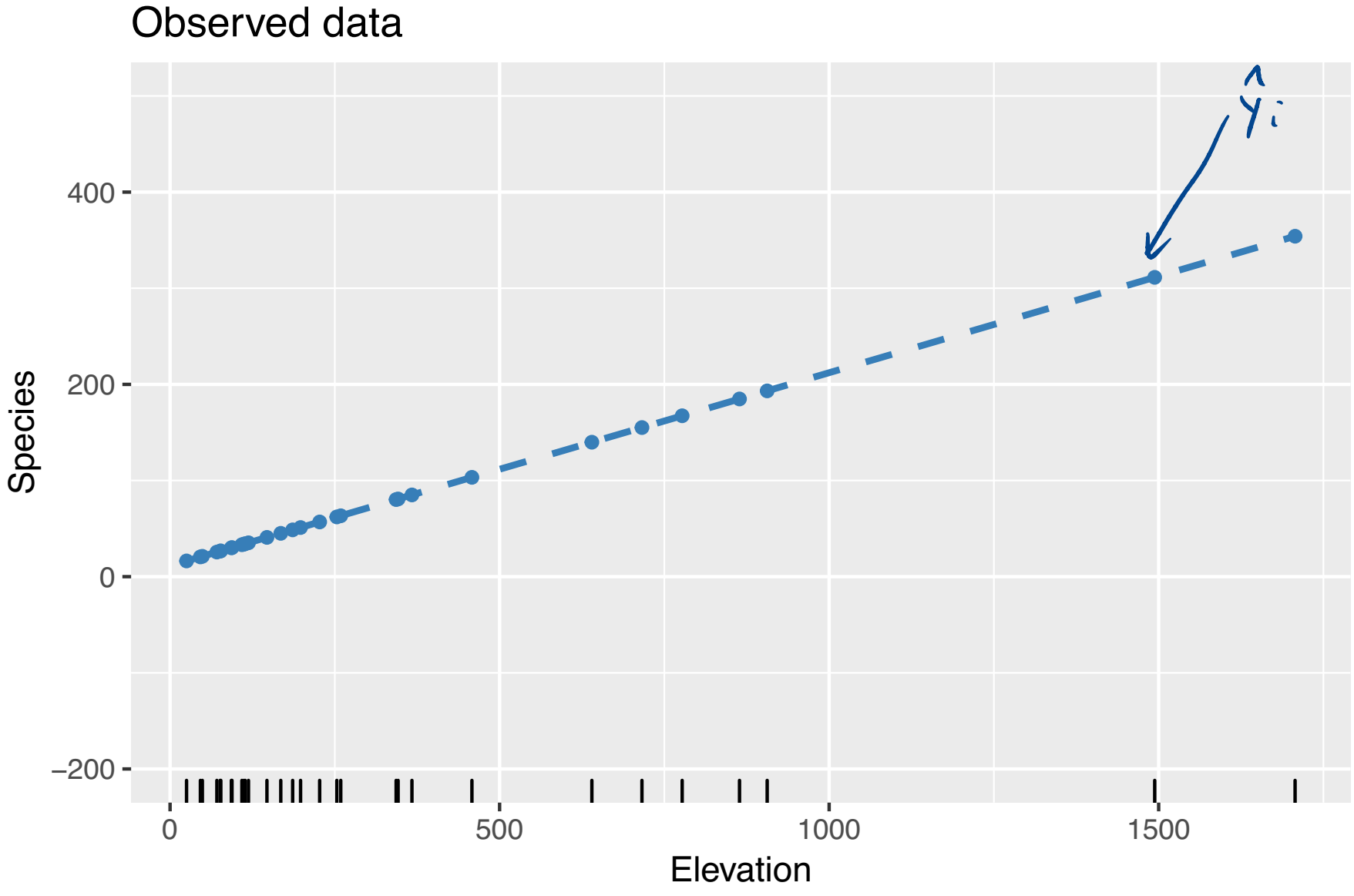
(I'll illustrate with simple linear regression, Faraway does multiple case in 3.6)



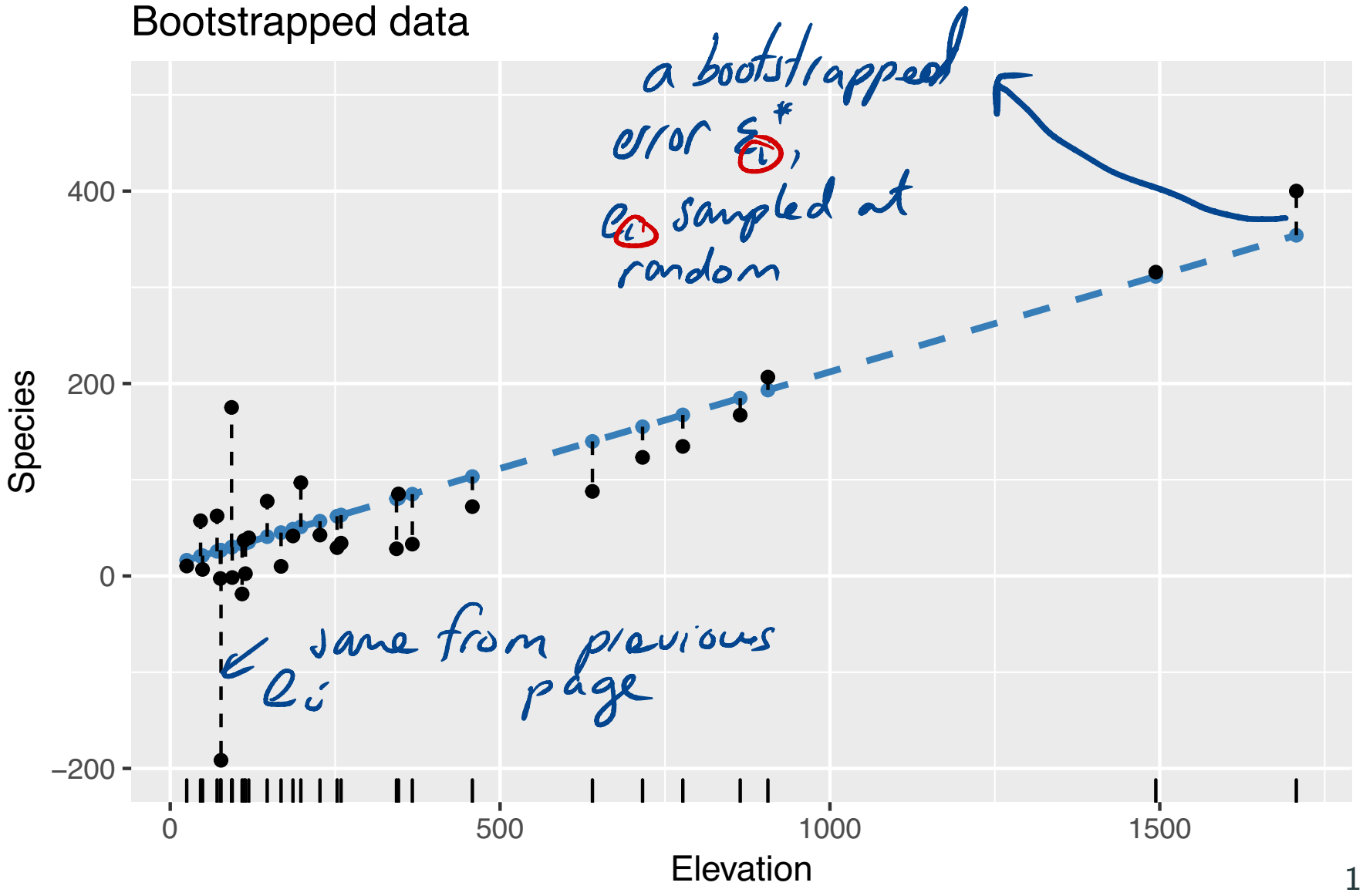
Bootstrap: 1. Find $\hat{\beta}$, $\hat{y}$, and e_i.



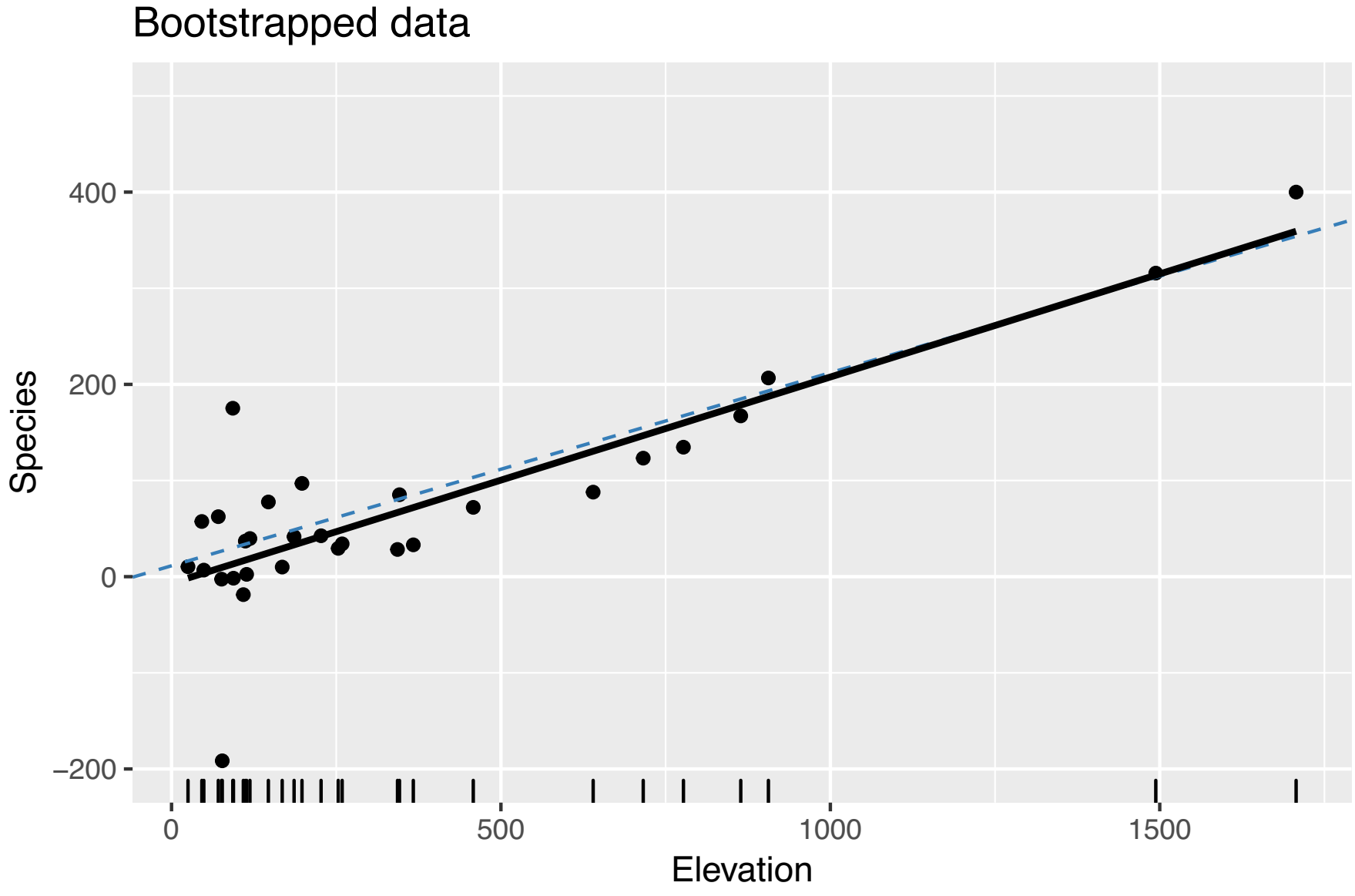
Bootstrap: Using fixed X, $\hat{beta}$ from observed data



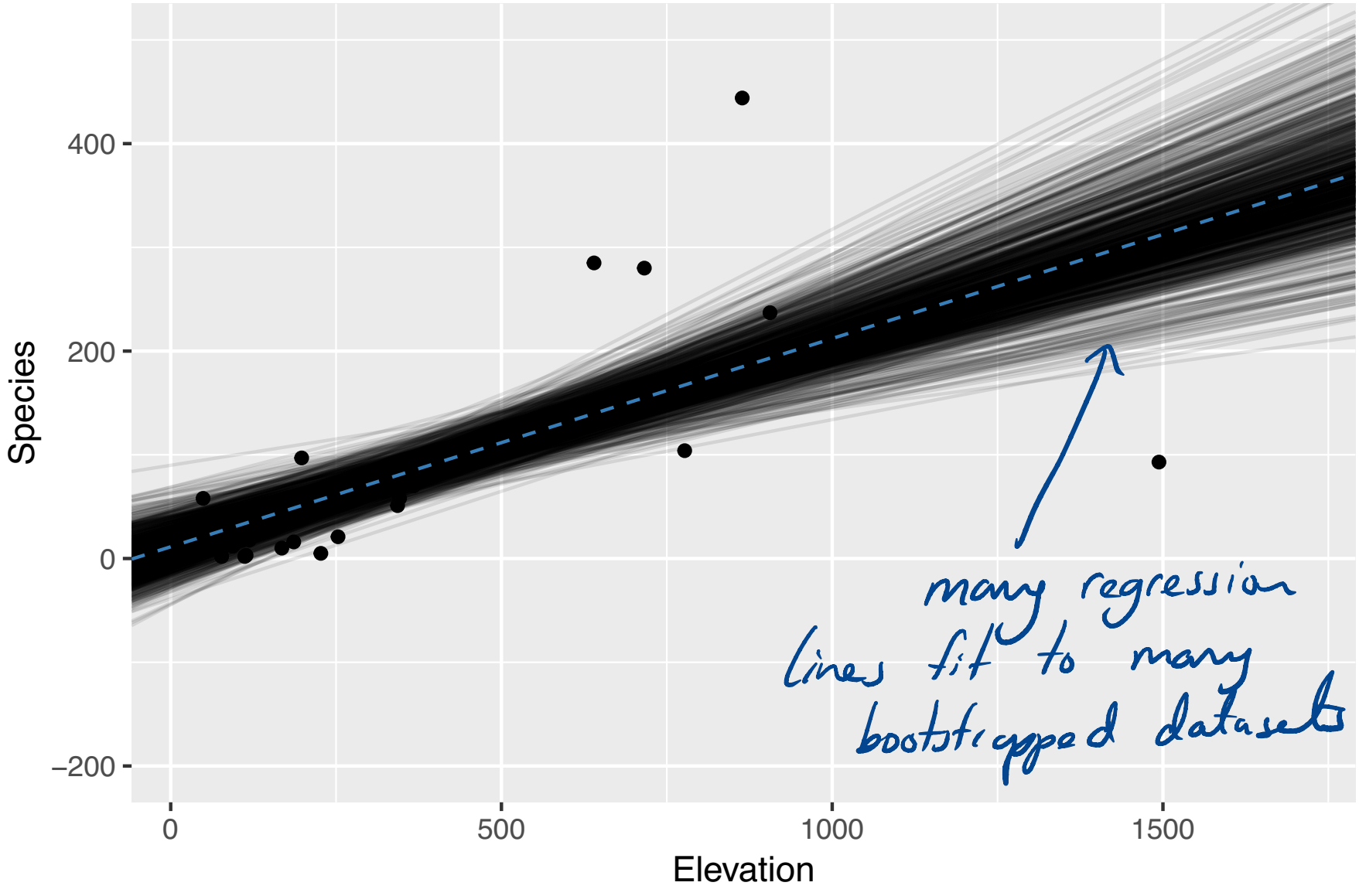
Bootstrap: 2. Resample residuals to construct bootstrapped response



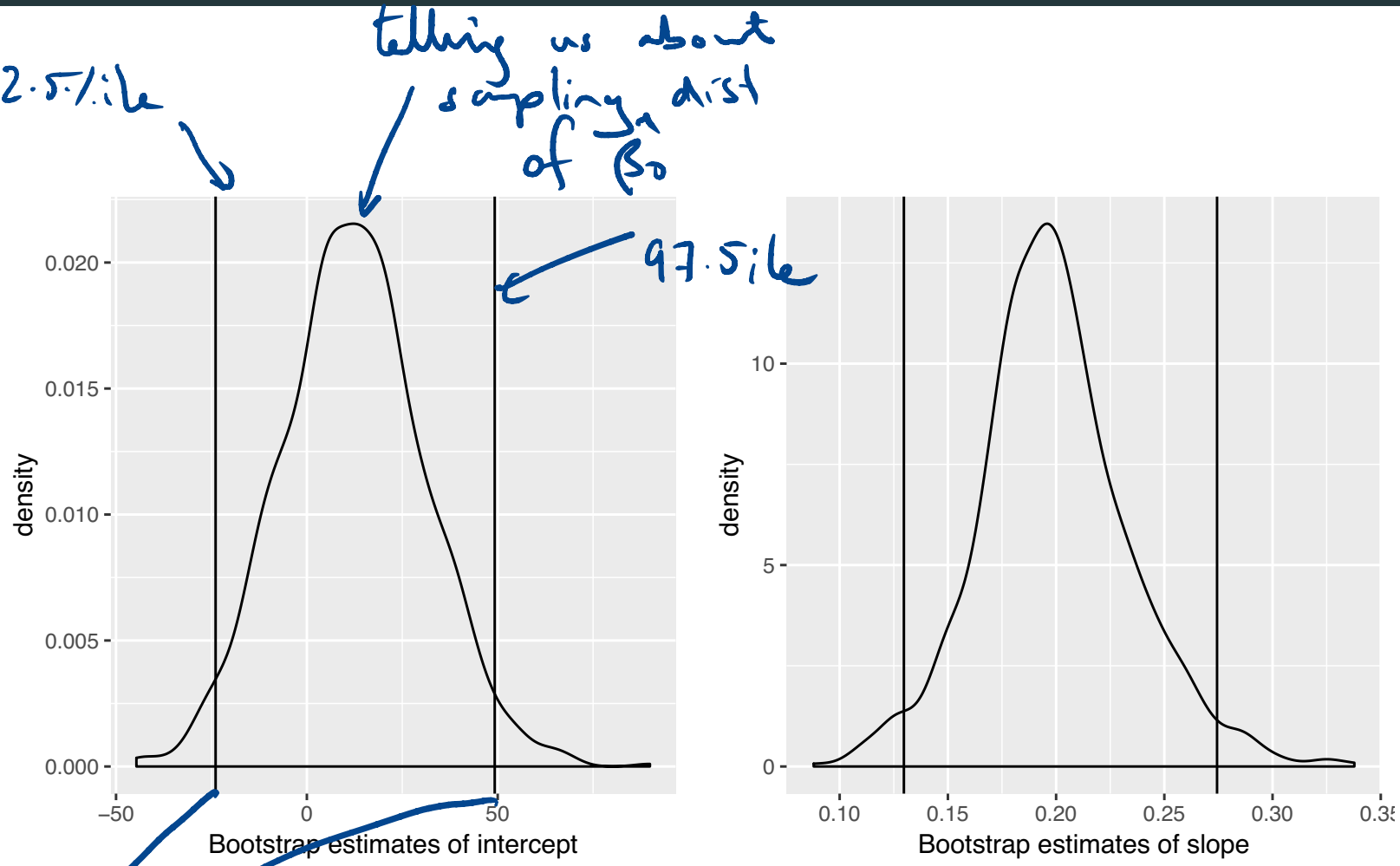
Bootstrap: 3. Fit regression model to bootstrapped response



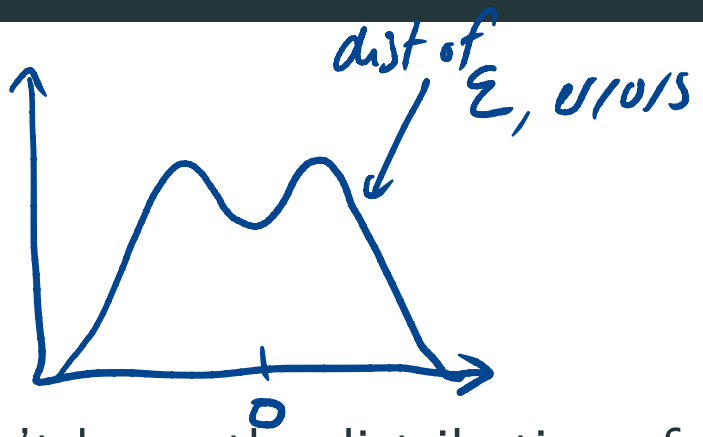
Bootstrap: 3. Repeat #2. and #3. many times

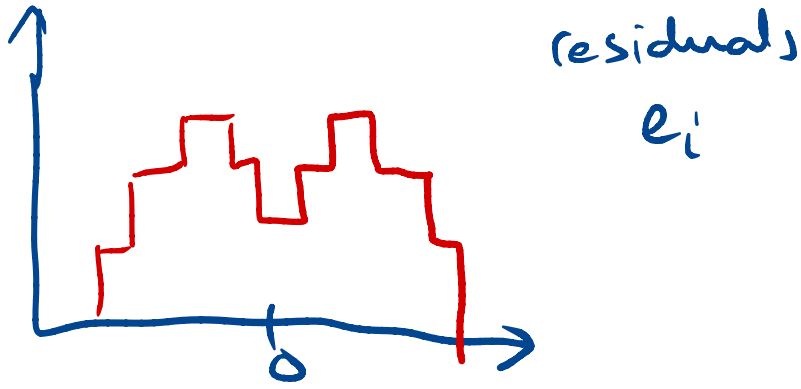


Examine distribution of estimates



High level: bootstrap idea





We don't know the distribution of the errors, but our best guess is probably the empirical c.d.f on the residuals.

Sampling from a random variable with a c.d.f. defined as the empirical c.d.f. of the residuals, boils down to sampling with replacement from residuals.

Limitations

We might rely on bootstrap confidence intervals when we are worried about the assumption of Normal errors. But, there are limitations.

- We still rely on the assumption that the errors are independent and identically distributed.
- Generally scaled residuals are used (residuals don't have the same variance, more later)
- An alternative bootstrap resamples the $(y_i, x_{i1}, \ldots, x_{ip})$ vectors, i.e. resamples the rows of the data, a.k.a *resampling cases bootstrap*.

↳ choice should depend on study design.

Experiment: resample residuals Sampling study:

Limitations

We might rely on bootstrap confidence intervals when we are worried about the assumption of Normal errors. But, there are limitations.

- We still rely on the assumption that the errors are independent and identically distributed.
- Generally scaled residuals are used (residuals don't have the same variance, more later)
- An alternative bootstrap resamples the $(y_i, x_{i1}, \ldots, x_{ip})$ vectors, i.e. resamples the rows of the data, a.k.a *resampling cases bootstrap*.