Bootstrap CIs

ST552 Lecture 13

Charlotte Wickham

Feb 08, 2016

Today

- Finish causal inference
- Bootstrap intervals

Bootstrap confidence intervals

What if ϵ are not from a Normal distribution?
The central limit theorem kicks in, so with large samples, even when the errors aren't Normal,

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

The bootstrap is one approach to estimate the sampling distribution of $\hat{\beta}$.

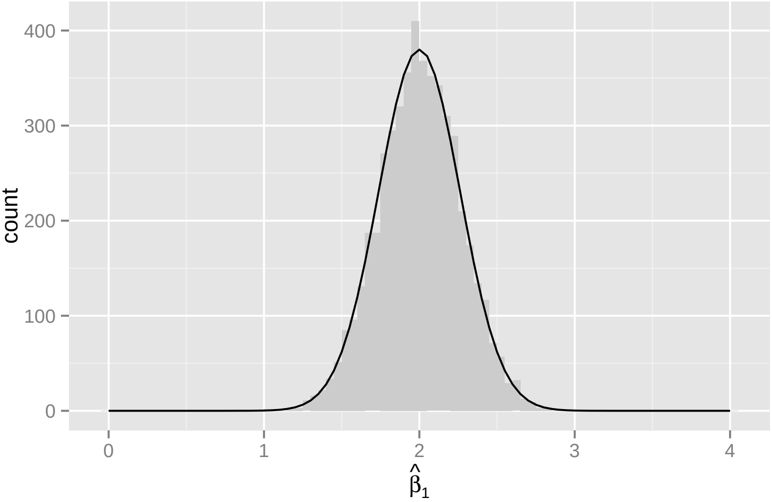
Outline

- What do we do if we know everything? Simulation.
- How does the bootstrap approximate that process?
- In practice
- Limitations

Simulation

To understand the sampling distribution of $\hat{\beta}$ we could use simulation.
Just like in HW#4. We know β and the distribution of ϵ.

1. Fix X
2. For $k = 1, \ldots, B$
    2.1 Generate errors, $\epsilon_i \overset{i.i.d}{\sim} Normal(0, \sigma^2)$
    2.2 Construct y, using the model, $y = X\beta + \epsilon$
    2.3 Use least squares to find $\hat{\beta}^*_{(k)}$
3. Examine the distribution of $\hat{\beta}^*$ and compare to β.

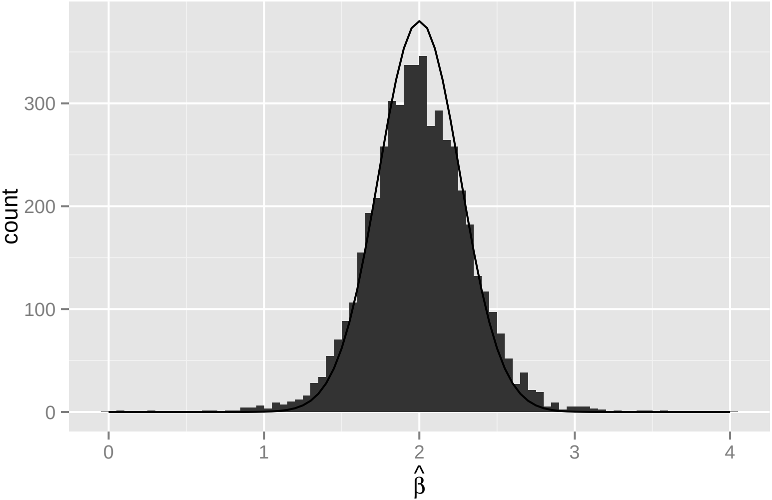



```
> quantile(ests$X1, c(0.025, 0.975))
    2.5%    97.5%
1.455147 2.539186
```

Simulation

If we want to know what happens to the distribution of $\hat{\beta}$ when the errors aren't Normal, we could assume some distribution for them and use simulation.
So, swap out step 2.1 for some other distribution. Let's say, Student's t with 3 d.f.
Just like in HW#4. We know β and the distribution of ϵ.

1. Fix X
2. For $k = 1, \ldots, B$
    2.1 Generate errors, $\epsilon_i \overset{i.i.d}{\sim}$ Student's-t$_3$
    2.2 Construct y, using the model, $y = X\beta + \epsilon$
    2.3 Use least squares to find $\hat{\beta}^*_{(k)}$
3. Examine the distribution of $\hat{\beta}^*$ and compare to β




```
> quantile(ests_t$X1, c(0.025, 0.975))
    2.5%     97.5%
1.355579 2.631276
```

Bootstrapping regression

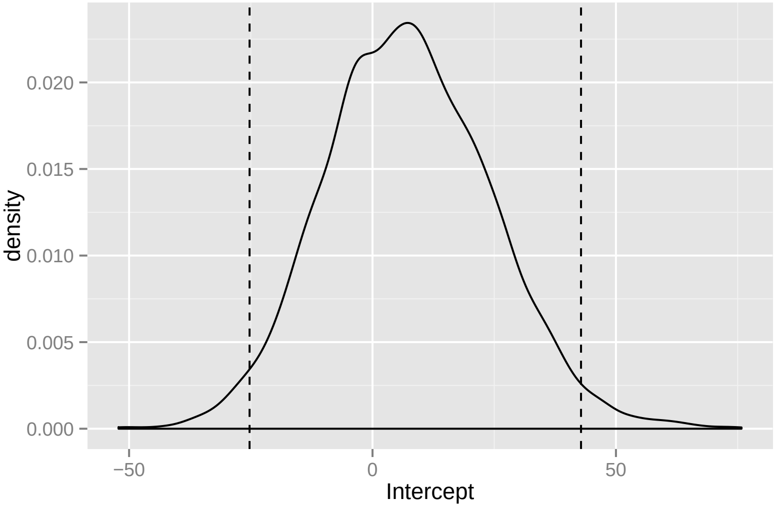
In a real life application we don't know β or the actual distribution of the errors. But we have some reasonable guesses we could make.

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

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

A naive confidence interval for β_j is the 2.5% and 97.5% quantiles of the distribution of $\hat{\beta}^*$. *(This relies on E$\left(\hat{\beta}^*\right) = \hat{\beta}$, and there are better methods)*

Example - Faraway



A reminder of the bootstrap idea

We don't know the distribution of some random variable Z but we can estimate it with observations of the random variable
$Z_i, \quad i = 1, \ldots, n.$
Usually, we think about this as using the empirical c.d.f. of Z_i to approximate the true c.d.f. of Z.
In practice, sampling from a random variable with a c.d.f. defined as the emprical c.d.f. of a set of numbers, Z_i, boils down to sampling with replacement from Z_i.

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 indentically 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.