The Bootstrap

Nate Wells

Math 243: Stat Learning

October 16th, 2020

Outline

In today's class, we will...

- Investigate the Bias-Variance trade-off
- Discuss the bootstrap for estimating variance of error

Section 1

The Bias-Variance Trade-off

Example

See .html and .Rmd file on course webpage for live-coded notes

Section 2

The Bootstrap

Why Bootstrap?

So, you want to know how a particular statistic is distributed?

Why Bootstrap?

So, you want to know how a particular statistic is distributed?

- Suppose you are interested in the distribution of slopes $\hat{\beta}_1$ in an SLR under random sampling:

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

Why Bootstrap?

So, you want to know how a particular statistic is distributed?

- Suppose you are interested in the distribution of slopes $\hat{\beta}_1$ in an SLR under random sampling:

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

- The classic approach:

Why Bootstrap?

So, you want to know how a particular statistic is distributed?

- Suppose you are interested in the distribution of slopes $\hat{\beta}_1$ in an SLR under random sampling:

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

- The classic approach:
    - Write the statistic $\hat{\beta}_1$ as a function of the random observations $x_1, \cdot, x_n$ and use properties of random variables to derive the theoretical distribution. Make some (sometimes unreasable) simplifying assumptions

Why Bootstrap?

So, you want to know how a particular statistic is distributed?

- Suppose you are interested in the distribution of slopes $\hat{\beta}_1$ in an SLR under random sampling:

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

- The classic approach:
  - Write the statistic $\hat{\beta}_1$ as a function of the random observations $x_1, \cdot, x_n$ and use properties of random variables to derive the theoretical distribution. Make some (sometimes unreasable) simplifying assumptions
  - Look up the theoretical distribution based on someone else's attempt to do part (1).

Why Bootstrap?

So, you want to know how a particular statistic is distributed?

- Suppose you are interested in the distribution of slopes $\hat{\beta}_1$ in an SLR under random sampling:

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

- The classic approach:
  - Write the statistic $\hat{\beta}_1$ as a function of the random observations $x_1, \cdot, x_n$ and use properties of random variables to derive the theoretical distribution. Make some (sometimes unreasable) simplifying assumptions
  - Look up the theoretical distribution based on someone else's attempt to do part (1).
  - Hope that the sample size is large enough to allow the Central Limit Theorem to come into play so that the statistic is approximately Normal

The Resampling Approach

As an alternative to using the theoretical distribution, use simulation to approximate.

The Resampling Approach

As an alternative to using the theoretical distribution, use simulation to approximate.

- The optimistic approach:

The Resampling Approach

As an alternative to using the theoretical distribution, use simulation to approximate.

- The optimistic approach:
    - Generate a large number of samples and compute the statistic of interest on each

The Resampling Approach

As an alternative to using the theoretical distribution, use simulation to approximate.

- The optimistic approach:
    - Generate a large number of samples and compute the statistic of interest on each
    - Plot and summarize the distribution of the statistic.

The Resampling Approach

As an alternative to using the theoretical distribution, use simulation to approximate.

- The optimistic approach:
  - Generate a large number of samples and compute the statistic of interest on each
  - Plot and summarize the distribution of the statistic.
  - The problem?

The Resampling Approach

As an alternative to using the theoretical distribution, use simulation to approximate.

- The optimistic approach:
    - Generate a large number of samples and compute the statistic of interest on each
    - Plot and summarize the distribution of the statistic.
    - The problem?
- The bootstrap approach:

The Resampling Approach

As an alternative to using the theoretical distribution, use simulation to approximate.

- The optimistic approach:
  - Generate a large number of samples and compute the statistic of interest on each
  - Plot and summarize the distribution of the statistic.
  - The problem?
- The bootstrap approach:
  - Assume that your sample is large enough to be "representative" of your population.

The Resampling Approach

As an alternative to using the theoretical distribution, use simulation to approximate.

- The optimistic approach:
  - Generate a large number of samples and compute the statistic of interest on each
  - Plot and summarize the distribution of the statistic.
  - The problem?
- The bootstrap approach:
  - Assume that your sample is large enough to be "representative" of your population.
  - Create a new bootstrap sample by sampling **with replacement** from your original sample, a number of times equal to your original sample size.

The Resampling Approach

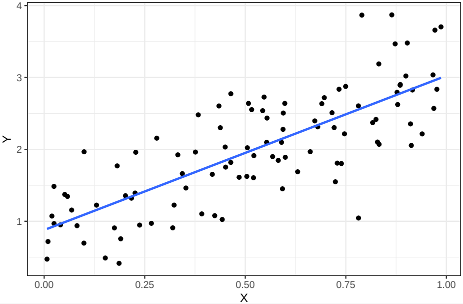
As an alternative to using the theoretical distribution, use simulation to approximate.

- The optimistic approach:
  - Generate a large number of samples and compute the statistic of interest on each
  - Plot and summarize the distribution of the statistic.
  - The problem?
- The bootstrap approach:
  - Assume that your sample is large enough to be "representative" of your population.
  - Create a new bootstrap sample by sampling **with replacement** from your original sample, a number of times equal to your original sample size.
  - Repeat the process to create many bootstrap samples. Compute the statistic of interest on each and plot the results.

Bootstrap Demo

Suppose $Y = 1 + 2 \cdot X + \epsilon$ with $\epsilon \sim N(0, 0.25)$.

```
set.seed(10101)
n<-100
X<-runif(n, 0, 1)
e<-rnorm(n, 0 ,.5)
Y<-1 + 2*X + e
d<-data.frame(X,Y)
```




```
my_mod<-lm(Y ~ X, data = d)
b1<-summary(my_mod)$coefficients[2,1]
b1
```

```
## [1] 2.146208
```

The Simulation Approach

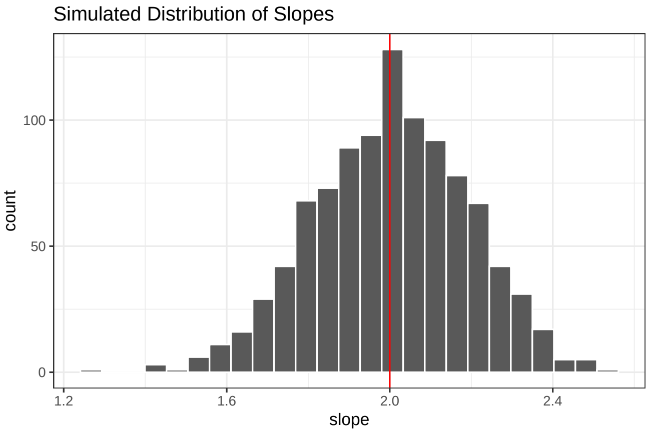
```r
set.seed(234)
trials<-1000 #Number of simulations
n<-100 #Number points in each simulation
X<-runif(n, 0, 1) # Generate random X; same for all sims
slopes<-data.frame() #Create empty dataframe for the slopes

for (i in 1:trials){
sim_e<-rnorm(n, 0 ,.5)
sim_Y<-1 + 2*X + sim_e
sim_d<-data.frame(X,sim_Y)
sim_mod<-lm(sim_Y ~ X, data = sim_d)
slopes<-rbind( slopes,
               data.frame(slope = summary(sim_mod)$coefficients[2,1]))
}
```

```r
head(slopes)
```

```
##      slope
## 1 2.238124
## 2 2.169395
## 3 1.904632
## 4 1.822680
## 5 1.846352
## 6 2.042824
```

Simulation Distribution

```
ggplot(slopes, aes(x = slope))+
  geom_histogram(bins= 25, color = "white")+theme_bw()+
  labs(title = "Simulated Distribution of Slopes")+
  geom_vline(xintercept = 2, color = "red")
```



The Bootstrap Approach

We have 1 sample:

```
head(d)
```

```
##           X         Y
## 1 0.1903066 0.7556851
## 2 0.9108393 2.3541632
## 3 0.2277161 1.9598872
## 4 0.8249905 2.4167019
## 5 0.9155760 2.8261117
## 6 0.5052083 2.0218132
```

But can create a bootstrap sample:

```
set.seed(135)
a_bootstrap_sample<-sample_n(d, size = n, replace = T)
```

The Bootstrap Approach

We have 1 sample:

```
head(d)
```

```
##           X         Y
## 1 0.1903066 0.7556851
## 2 0.9108393 2.3541632
## 3 0.2277161 1.9598872
## 4 0.8249905 2.4167019
## 5 0.9155760 2.8261117
## 6 0.5052083 2.0218132
```

But can create a bootstrap sample:

```
set.seed(135)
a_bootstrap_sample<-sample_n(d, size = n, replace = T)
```

Duplicates?

```
common<-intersect(a_bootstrap_sample, d)
length(common$X)
```

```
## [1] 66
```

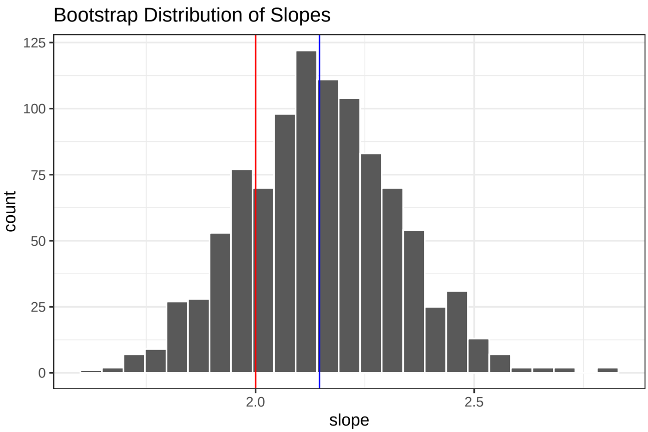
The Bootstrap Approach, cont'd

Now, we create 1000 bootstraps and calculate the slope of each

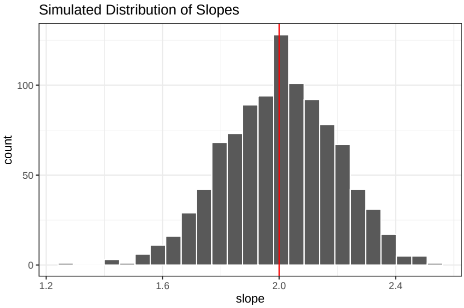
```
trials<-1000
bootstraps<-data.frame()
for (i in 1:trials){
  boot<-sample_n(d, size = n, replace = T)
  my_mod<-lm(Y ~ X , data = boot)
  bootstraps<- rbind(bootstraps,
                     data.frame(slope = summary(my_mod)$coefficients[2,1]))
}
```

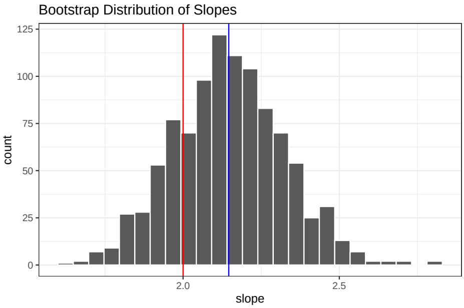
Bootstrap Distribution

```
ggplot(bootstraps, aes(x = slope))+
  geom_histogram(bins= 25, color = "white")+theme_bw() +
  labs(title = "Bootstrap Distribution of Slopes") +
  geom_vline(xintercept = b1, color = "blue" )+
  geom_vline(xintercept = 2, color = "red")
```

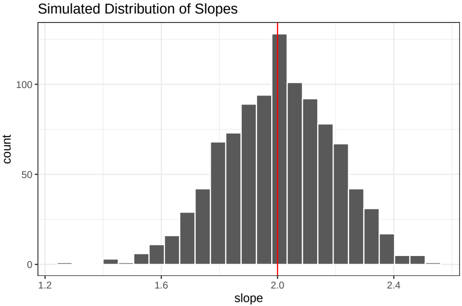


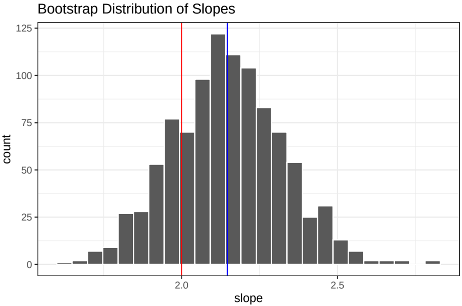
Side-by-Side Comparison





Side-by-Side Comparison





How does this related to the decomposition of MSE into Bias and Variance?

CV verus Bootstrapping

Both are computationally intensive methods that involve sampling from your data set to learn more about your estimate/model.

CV verus Bootstrapping

Both are computationally intensive methods that involve sampling from your data set to learn more about your estimate/model.

Cross-validation: Often used for *model assessment* and *model selection*.

CV verus Bootstrapping

Both are computationally intensive methods that involve sampling from your data set to learn more about your estimate/model.

Cross-validation: Often used for *model assessment* and *model selection*.

- Partition data into test and train
- Fit model to train, predict on test
- Iterate though all possible *folds*
- Compute aggregate measure of predictive ability

CV verus Bootstrapping

Both are computationally intensive methods that involve sampling from your data set to learn more about your estimate/model.

Cross-validation: Often used for *model assessment* and *model selection*.

- Partition data into test and train
- Fit model to train, predict on test
- Iterate though all possible *folds*
- Compute aggregate measure of predictive ability

Bootstrapping: Often used for *quantifying uncertainty*.

CV verus Bootstrapping

Both are computationally intensive methods that involve sampling from your data set to learn more about your estimate/model.

Cross-validation: Often used for *model assessment* and *model selection*.

- Partition data into test and train
- Fit model to train, predict on test
- Iterate though all possible *folds*
- Compute aggregate measure of predictive ability

Bootstrapping: Often used for *quantifying uncertainty*.

- Draw a bootstrap sample of size n from your data *with replacement*.
- Compute estimate of interest
- Consider distribution of bootstrap estimates over many samples