Bootstrap Inference

A brief introduction using logistic regression

Brook Luers
Department of Statistics, University of Michigan
luers@umich.edu
brookluers.com

March 26, 2020

Overview

We are estimating a parameter θ using a statistic $\hat{\theta}$.

The bootstrap:

- provides confidence intervals, hypothesis tests for θ
- without restrictive assumptions
- uses a computer to approximate the sampling distribution of $\hat{\theta}$

Especially useful when:

- difficult to mathematically derive
  - the variance of $\hat{\theta}$
  - the sampling distribution of $\hat{\theta}$
- modeling assumptions are suspect

Example: no bootstrap needed

$(X_1, Y_1) \ldots, (X_n, Y_n)$ are independent observations and

$$Y_i \mid X_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2).$$

Estimate the slope $\theta = \beta_1$ using simple linear regression.

If $\hat{\beta}_1$ is the simple linear regression slope, we can show that

$$\hat{\theta} = \hat{\beta}_1 \sim N\left(\beta_1, \sigma^2 / \sum_i (X_i - \overline{X})^2\right),$$

and we can use this fact to compute a confidence interval.

In many situations, it is not as easy to derive the exact sampling distribution of $\hat{\theta}$.

Example: The Challenger disaster

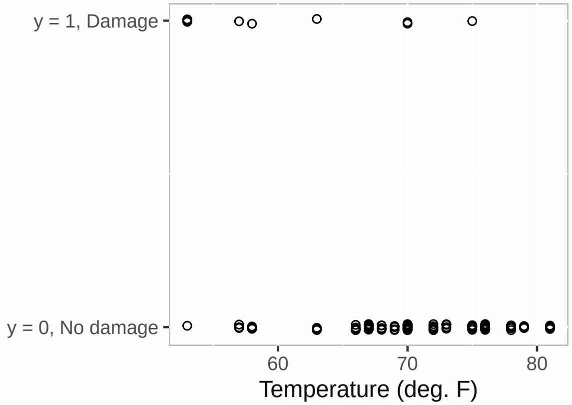
On January 28, 1986, the Space Shuttle Challenger disintigrated after 73 seconds of flight.

This was caused by failures in rocket booster parts called o-rings.

Prior to launch, there were concerns about the effect of low temperatures on o-ring performance.

Example: The Challenger disaster

Data on 138 o-rings used in previous launches:



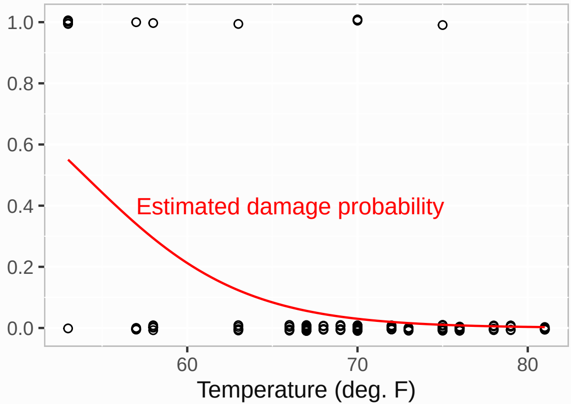
Example: The Challenger disaster

We fit a logistic regression model to relate damage probability to launch temperature:

$$\log\left(\frac{P(Y=1 \mid \texttt{temp})}{1-P(Y=1 \mid \texttt{temp})}\right) = \beta_0 + \beta_1 \texttt{temp}.$$

Example: The Challenger disaster

$$\log\left(\frac{P(Y=1 \mid \texttt{temp})}{1-P(Y=1 \mid \texttt{temp})}\right) = \beta_0 + \beta_1 \texttt{temp} \qquad \hat{\beta}_0 = 11.66, \hat{\beta}_1 = -0.22$$



Parameter of interest

Suppose we want to estimate θ, the lowest temperature at which the probability of o-ring damage is less than 20 percent.

We would like a confidence interval for θ based on our logistic regression model.

Parameter of interest

$\theta = \texttt{temp}$

$\tau = 0.2 =$ damage probability

Logistic regression model:

$$\log\left(\frac{P(Y=1 \mid \texttt{temp})}{1 - P(Y=1 \mid \texttt{temp})}\right) = \beta_0 + \beta_1 \texttt{temp}$$

$$\log\left(\frac{\tau}{1-\tau}\right) = \beta_0 + \beta_1 \theta$$

$$\theta = \frac{1}{\beta_1}\left(\log\frac{\tau}{1-\tau} - \beta_0\right)$$

Plug-in estimate of θ

Based on our logistic regression model:

$$\log\left(\frac{P(Y=1 \mid \texttt{temp})}{1-P(Y=1 \mid \texttt{temp})}\right) = \beta_0 + \beta_1 \texttt{temp} \qquad \hat{\beta}_0 = 11.66, \hat{\beta}_1 = -0.22$$

$$\hat{\theta} = \frac{1}{\hat{\beta}_1}\left(\log \frac{0.2}{1-0.2} - \hat{\beta}_0\right) = 60.3$$

Estimated minimum launch temperature: 60 degrees
(to ensure damage probability of less than 0.2)

Confidence interval for θ?

Confidence interval for θ

$$\hat{\theta} = \frac{1}{\hat{\beta}_1}\left(\log \frac{0.2}{1-0.2} - \hat{\beta}_0\right) = 60.3$$

To compute a CI, we usually need to know the sampling distribution of $\hat{\theta}$, or at least its variance.

Given $\mathrm{Var}(\hat{\theta})$, we could form the approximate 95% CI

$$\hat{\theta} \pm 2\sqrt{\mathrm{Var}(\hat{\theta})}.$$

In this case, it is not clear how to derive $\mathrm{Var}(\hat{\theta})$.

With the bootstrap, we can compute a CI for θ **even if we do not know how to mathematically derive the variance of $\hat{\theta}$.**

The bootstrap: how it works

Given a random sample *S* of size *n* drawn from a population, we compute $\hat{\theta}$ to estimate θ.

If we could repeatedly sample from this population, we could repeatedly compute $\hat{\theta}$ and examine its sampling distribution.

The bootstrap treats the original sample *S* as a stand-in for the population.

Repeated sampling from *S* approximates repeated sampling from the population.

The bootstrap: how it works

Given a random sample S of size n drawn from a population, we compute $\hat{\theta}$ to estimate θ.

The bootstrap:

1. Generate B "bootstrap samples"
   - with n elements per sample
   - each element randomly drawn from S, with replacement
2. Compute the estimates $\hat{\theta}^*_b$ using the bth bootstrap sample, $b = 1, \ldots, B$.
3. The collection $\hat{\theta}^*_1, \ldots, \hat{\theta}^*_B$ approximates the sampling distribution of $\hat{\theta}$.

Bootstrap CI

The collection $\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*$ approximates the distribution of $\hat{\theta}$.

Based on this principle,

$$\mathrm{SE}_{\mathrm{boot}}(\hat{\theta}) = \sqrt{\frac{1}{B-1}\sum_b (\hat{\theta}_b^* - \overline{\theta^*})^2}$$

is a reasonable estimate of $\sqrt{\mathrm{Var}(\hat{\theta})}$.

Given $\sqrt{\mathrm{Var}(\hat{\theta})}$, an approximate 95% CI is

$$\hat{\theta} \pm 2 \times \sqrt{\mathrm{Var}(\hat{\theta})}.$$

A bootstrap CI replaces $\sqrt{\mathrm{Var}(\hat{\theta})}$ with its bootstrap estimate:

$$\hat{\theta} \pm 2 \times \mathrm{SE}_{\mathrm{boot}}(\hat{\theta}).$$

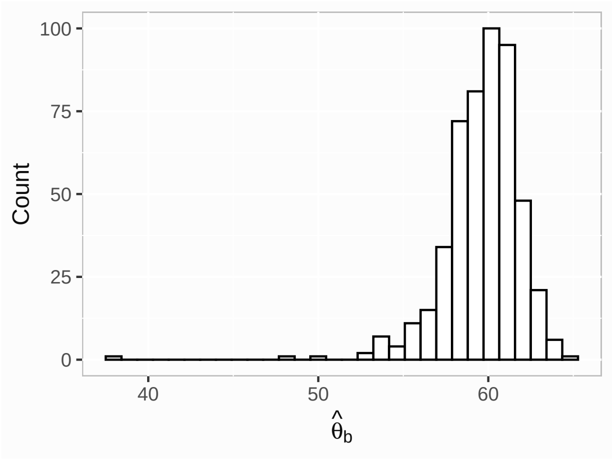
Bootstrap sampling

Partial R code

```r
# A vector to store the resulting hat(theta)_b
thetas_boot <- vector('numeric', B)

for (b in 1:B){

  # random indices for sample b
  ix_b <- sample(1:n, size=n, replace = TRUE)

  # bootstrap sample b
  data_b <- data_original[ix_b, ]

  # fit the regression model using the bootstrap sample
  model_b <- glm(fail ~ temp, family=binomial, data = data_b)

  b0 <- coef(model_b)[1] # hat(beta)_0
  b1 <- coef(model_b)[2] # hat(beta)_1

  # compute hat(theta)_b
  thetas_boot[b] <- (log(0.2 / (1 - 0.2)) - b0) / b1
}
```

Bootstrap samples

Histogram of $\hat{\theta}_b^*$ approximates the sampling distribution of $\hat{\theta}$:



Bootstrap CI

Original estimate $\hat{\theta}$:

```
theta_hat

## [1] 60.34807
```

Bootstrap standard error:

```
(se_hat_boot <- sd(thetas_boot))

## [1] 2.295753
```

An approximate 95% confidence interval:

```
c(theta_hat - 2 * se_hat_boot,
  theta_hat + 2 * se_hat_boot)

## [1] 55.75656 64.93958
```

Summary

The bootstrap:

- provides standard errors, confidence intervals, hypothesis tests
- without restrictive assumptions
- useful when mathematical analysis is not possible
- can be computationally expensive

Further reading:

- Bradley Efron and Robert Tibshirani. An introduction to the bootstrap. CRC press, 1994.
- Bryan Manly. Randomization, bootstrap and Monte Carlo methods in biology. CRC press, 2007.
- Christopher Mooney and Robert Duval. Bootstrapping: A nonparametric approach to statistical inference. Sage, 1993.

Thank you

Slides and code: github.com/brookluers/bootstrap-challenger