Lecture 6: Estimating Uncertainty

STATS 202: Data Mining and Analysis

Linh Tran

tranlm@stanford.edu

Department of Statistics
Stanford University

July 12, 2021

Announcements

- HW1 grades posted (solutions on Piazza).
- HW2 due Friday.
- Midterm is in 9 days.
  - Will be posted to Piazza/Gradescope at 4PM Wednesday and due in Gradescope within 24 hours.
  - Open book (ISL/ESL)
  - Let me know if you need special accomodations
  - Solutions to practice midterm will be posted on Wednesday
  - Review this Friday

Outline

- The bootstrap
  - Intro
  - Types, uses, etc.
  - Bagging
- The jackknife
  - Intro
  - Bootstrap vs jackknife

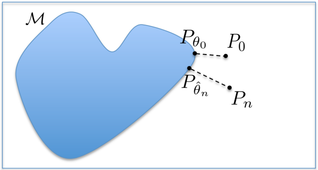
Recap

Previously, we:

- Defined data generating mechanisms as true functions
- Proposed methods of estimating the functions
- Covered ways of evaluating model performance

How precise are our estimates?

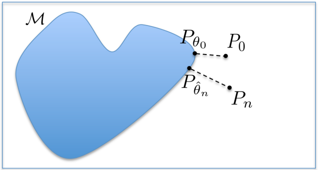
Standard errors



Recall:

- Using our data P_n, we can estimate our parameter ψ_0

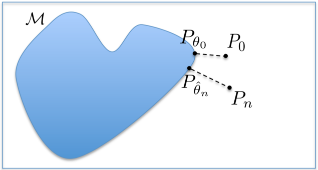
Standard errors



Recall:

- Using our data P_n, we can estimate our parameter ψ_0
- Because our data is random, the estimate $\hat{\psi}_n$ is random

Standard errors



Recall:

- Using our data P_n, we can estimate our parameter ψ_0
- Because our data is random, the estimate $\hat{\psi}_n$ is random
- If ψ_0 is e.g. a linear model coefficient, then can use closed form formulas, e.g.

$$\mathrm{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{1}$$

Standard errors

An example: Standard errors in linear regression

```
Residuals:
    Min      1Q  Median      3Q     Max
-15.594  -2.730  -0.518   1.777  26.199

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.646e+01  5.103e+00   7.144 3.28e-12 ***
crim        -1.080e-01  3.286e-02  -3.287 0.001087 **
zn           4.642e-02  1.373e-02   3.382 0.000778 ***
indus        2.056e-02  6.150e-02   0.334 0.738288
chas         2.687e+00  8.616e-01   3.118 0.001925 **
nox         -1.777e+01  3.820e+00  -4.651 4.25e-06 ***
rm           3.810e+00  4.179e-01   9.116  < 2e-16 ***
age          6.922e-04  1.321e-02   0.052 0.958229
dis         -1.476e+00  1.995e-01  -7.398 6.01e-13 ***
rad          3.060e-01  6.635e-02   4.613 5.07e-06 ***
tax         -1.233e-02  3.761e-03  -3.280 0.001112 **
ptratio     -9.527e-01  1.308e-01  -7.283 1.31e-12 ***
black        9.312e-03  2.686e-03   3.467 0.000573 ***
lstat       -5.248e-01  5.072e-02 -10.347  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.745 on 492 degrees of freedom
Multiple R-Squared: 0.7406,     Adjusted R-squared: 0.7338
F-statistic: 108.1 on 13 and 492 DF,  p-value: < 2.2e-16
```

Standard errors

More generally: Obtain estimator's *sampling distribution*

Standard errors

More generally: Obtain estimator's *sampling distribution*

Example: The variance of a sample $x_1, x_2, ..., x_n$

$$\hat{\sigma}_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{2}$$

Standard errors

More generally: Obtain estimator's *sampling distribution*

Example: The variance of a sample $x_1, x_2, \ldots, x_n$

$$\hat{\sigma}_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{2}$$

How to get the standard error of $\hat{\sigma}_n^2$

1. Assume $x_1, x_2, \ldots, x_n \overset{iid}{\sim} \mathcal{N}(\mu_0, \sigma_0^2)$
2. Assume that $\hat{\sigma}_n^2$ is close to σ_0^2 and $\bar{x}$ is close to μ_0
3. Then $\hat{\sigma}_n^2(n-1)$ has been shown to have a χ-squared distribution with n degrees of freedom
4. The SD of this sampling distribution is the standard error

What if:

- The sampling distribution is not easy to derive?
- Our distributional assumptions break down?

Standard errors

What if:

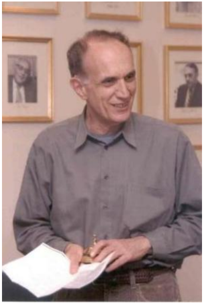
- The sampling distribution is not easy to derive?
- Our distributional assumptions break down?

Some possible options:

1. Bootstrap
2. Jackknife
3. Influence functions
    - Beyond scope of this course

The Bootstrap

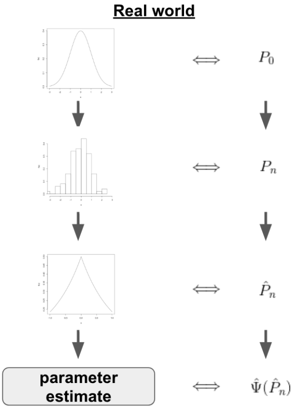
Method to simulate generating from the true distribution P_0



- Provides standard error of estimates
- Popularized by Brad Efron (Stanford)
  - Wrote "An Introduction to the Bootstrap" with Robert Tibshirani
- Very popular among practitioners
- Computer intensive (d/t the approach)

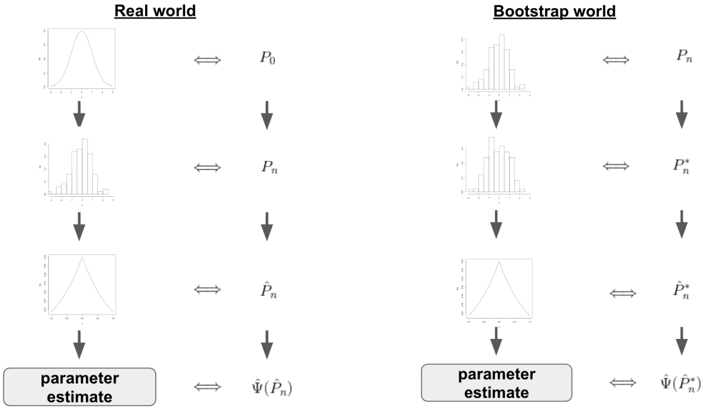
The bootstrap

Method to simulate generating from the true distribution P_0

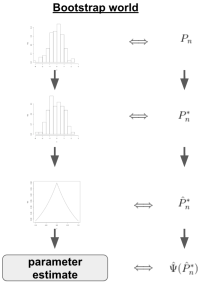


The bootstrap

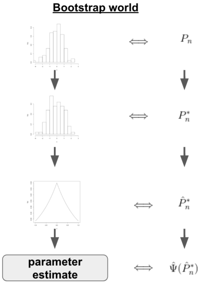
Method to simulate generating from the true distribution P_0



The bootstrap

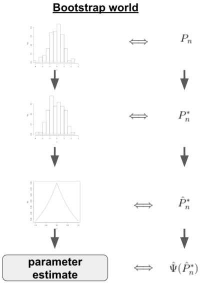


- This resampling method is repeated (say, B times) until we have "*enough*" iterations to get a stable distribution.
  - Results in a simulated sampling distribution



- This resampling method is repeated (say, B times) until we have "*enough*" iterations to get a stable distribution.
  - Results in a simulated sampling distribution
- The SD of this sampling distribution is our estimated standard error

The bootstrap

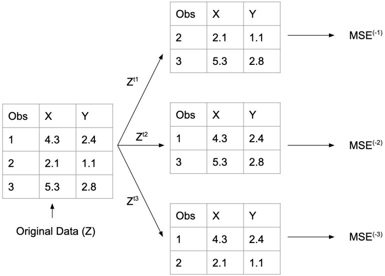


- This resampling method is repeated (say, B times) until we have "*enough*" iterations to get a stable distribution.
  - Results in a simulated sampling distribution
- The SD of this sampling distribution is our estimated standard error
- n.b. Two approximations are made:

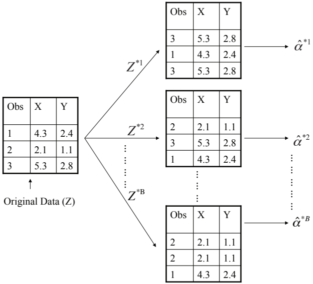
$$SE(\hat{\psi}_n)^2 \overbrace{\approx}^{\text{not so small}} \hat{SE}(\hat{\psi}_n)^2 \overbrace{\approx}^{\text{small}} \hat{SE}_B(\hat{\psi}_n)^2 \tag{3}$$

Bootstrap vs Cross-validation

Cross-validation: provides estimates of the (test) error.



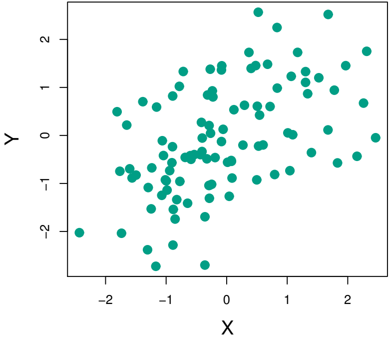
Bootstrap: provides the (standard) error of estimates.



Example. Investing in two assets

Suppose that X and Y are the returns of two assets.

The returns are observed every day, i.e. $(x_1, y_1), \ldots, (x_n, y_n)$.



Example. Investing in two assets

We only have a fixed amount of money to invest, so we'll invest

- α in X and $(1-\alpha)$ in Y, where α is between 0 and 1, i.e.
$$\alpha X + (1-\alpha) Y \tag{4}$$

Example. Investing in two assets

We only have a fixed amount of money to invest, so we'll invest

- α in X and $(1-\alpha)$ in Y, where α is between 0 and 1, i.e.

$$\alpha X + (1-\alpha)Y \tag{4}$$

Our goal: Minimize the variance of our return as a function of α

- One can show that the optimal α_0 is:

$$\alpha_0 = \frac{\sigma_Y^2 - \sigma_{XY}}{\sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY}} \tag{5}$$

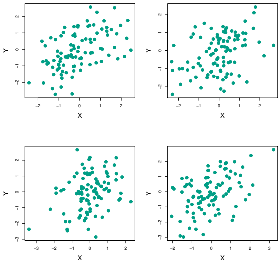
- which we can estimate using our data, i.e.

$$\hat{\alpha}_n = \frac{\hat{\sigma}_{Y,n}^2 - \hat{\sigma}_{XY,n}}{\hat{\sigma}_{X,n}^2 + \hat{\sigma}_{Y,n}^2 - 2\hat{\sigma}_{XY,n}} \tag{6}$$

Example. Investing in two assets

If: we knew P_0, we could just resample the n observations and re-calculate $\hat{\alpha}_n$.

- We could iterate on this until we have enough estimates to form a sampling distribution
- Would then estimate the SE via the SD of the distribution

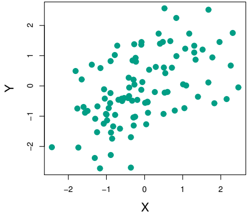


Four draws from P_0.

Example. Investing in two assets

Reality: We don't know P_0 and only have n observations.

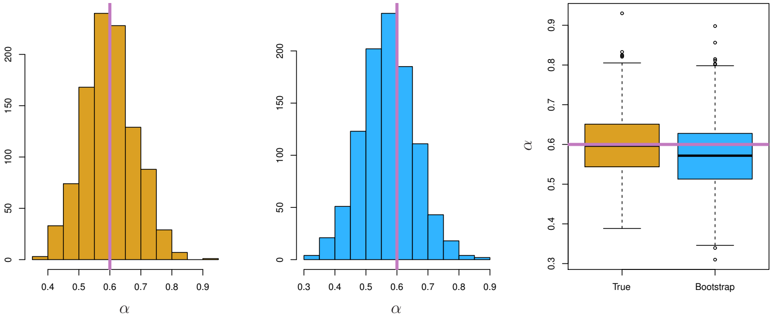
But: We can mimic as if we did know P_0.



- Assume that P_n is a good approximation of P_0
- Iteratively (say, B times):
  - Resample from P_n, i.e. sample from the n observations with replacement, n times (call this $P_n^{*,r}$)
  - Calculate $\hat{\alpha}_n$ from $P_n^{*,r}$ (call this $\hat{\alpha}_n^{*,r}$)
- Calculate the SD of the $\hat{\alpha}_n^{*,r}$ estimates, i.e.

$$\widehat{SE}_B(\hat{\alpha}_n) = \sqrt{\frac{1}{B-1}\sum_{r=1}^{B}\left(\hat{\alpha}_n^{*,r} - \frac{1}{B}\sum_{r'=1}^{B}\hat{\alpha}_n^{*,r'}\right)^2}$$

Bootstrap distribution vs true distribution



True (*left*) and bootstrap (*center*) sampling distributions

Bootstrap and error rates

Each bootstrap iteration will only have about 2/3 of the original data, i.e.

$$\mathbb{P}(x_j \notin P_n^b) = (1 - 1/n)^n \tag{8}$$

Bootstrap and error rates

Each bootstrap iteration will only have about 2/3 of the original data, i.e.

$$\mathbb{P}(x_j \notin P_n^b) = (1 - 1/n)^n \tag{8}$$

We could use the out of bag observations to calculate estimate our test set error, i.e.

$$\widehat{Err} = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{|C^{-i}|} \sum_{b \in C^{-i}} L(y_i, \hat{f}^{*b}(x_i)) \tag{9}$$

- Doing this still encounters 'training-set' bias (i.e. you're using less observations to estimate f_0).

Hypothetical Example. Patient headache

Let

- $X_{i,j}$ be an indicator that patient i took asprin on day j.
- $Y_{i,j}$ be an indicator that patient i had a headache on day j.

We want the standard error for the
$P(headache|asprinstatus)$

Hypothetical Example. Patient headache

Let

- $X_{i,j}$ be an indicator that patient i took asprin on day j.
- $Y_{i,j}$ be an indicator that patient i had a headache on day j.

We want the standard error for the
$P(headache|asprinstatus)$

Wrong way: Bootstrap over all i, j observations and calculate
$P(headache|asprin)$

Hypothetical Example. Patient headache

Let

- $X_{i,j}$ be an indicator that patient i took asprin on day j.
- $Y_{i,j}$ be an indicator that patient i had a headache on day j.

We want the standard error for the $P(headache|asprinstatus)$

Wrong way: Bootstrap over all i, j observations and calculate $P(headache|asprin)$

Right way: Bootstrap by patient id and calculate $P(headache|asprin)$

Example simulation

Let

$$Y_i, X_i \in \mathbb{R} : i = 1, 2, ..., n \ \ni \ Y_i = X_i + \epsilon_i : \epsilon_i \sim N(0, \sigma^2)$$

We wish to calculate the standard error of predictions.

Example simulation

Let
$Y_i, X_i \in \mathbb{R} : i = 1, 2, ..., n \ni \; Y_i = X_i + \epsilon_i : \epsilon_i \sim N(0, \sigma^2)$

We wish to calculate the standard error of predictions.

Method 1: Rely on asymptotic theory

$$\hat{se}(\hat{y}_i) = \sqrt{\hat{\sigma}^2 \left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{j=1}^{n} (x_j - \bar{x})^2} \right)} \tag{10}$$

Example simulation

Let

$Y_i, X_i \in \mathbb{R} : i = 1, 2, ..., n \ni \; Y_i = X_i + \epsilon_i : \epsilon_i \sim N(0, \sigma^2)$

We wish to calculate the standard error of predictions.

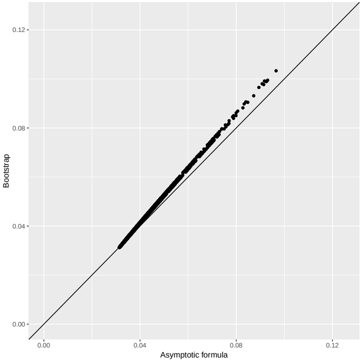
Method 1: Rely on asymptotic theory

$$\hat{se}(\hat{y}_i) = \sqrt{\hat{\sigma}^2 \left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{j=1}^{n} (x_j - \bar{x})^2} \right)} \tag{10}$$

Method 2: Bootstrap across B iterations and calculate

$$\hat{se}(\hat{y}_i) = \sqrt{\frac{1}{B-1} \sum_{b=1}^{B} (\hat{y}_i^b - \bar{y}_i^b)^2} \tag{11}$$

Example simulation



Comparison over $n = 1000$ observations

Bootstrap forms

Our presentation up to now has been on '*nonparameteric*' bootstrapping.

Intead, we could bootstrap the data other ways:

- **Parametric**: use the fitted model with some (e.g. Gaussian) noise to construct our resampled data.
- **Bayesian**: resample points using weights.
- **Residual**: resample errors and add to predictions.
- **Block**: resample blocks (accounting for correlations).
- etc...

Parametric Bootstrap

Let $X, Y \in \mathbb{R}$ and assume $Y_i = X_i + \epsilon_i : i = 1, 2, ..., n$.

Parametric Bootstrap:

$$Y_i^* = \hat{y}_i + \epsilon_i^*; \epsilon_i^* \sim N(0, \hat{\sigma}^2) : i = 1, 2, ..., n \tag{12}$$

Repeat B times and take standard deviation over the estimates.

Confidence intervals

Bootstrap standard errors can be used to compute confidence intervals, e.g.

- Normal-based interval
- Quantile interval
- Pivotal interval
- Studentized interval

Normal-based confidence interval

The same as calculating an interval under a normal distribution

- Switch out asymptotic standard error with bootstrap estimate
- Only works well if the distribution of the statistic is close to normal

Normal-based confidence interval

$$C_n = \hat{\psi}_n \pm z\alpha/2\hat{\text{se}}_{boot} \tag{13}$$

Quantile interval

Use the observed bootstrap distribution's quantiles, e.g. select 2.5% and 97.5% values.

- Can result in noticeably different estimates under skewed distributions.

Quantile confidence interval

$$C_n = \left(\hat{\psi}^*_{n,\alpha/2}, \hat{\psi}^*_{n,1-\alpha/2}\right) \tag{14}$$

Pivotal confidence interval

Let $R_n = R(X_1, \ldots, X_n, \psi_0)$ be a function who's distribution does not depend on ψ_0.

- We can construct a CI for R_n without knowing ψ_0
- Would then manipulate the CI to construct a CI for ψ_0
- AKA "*basic*" interval in R

Defining $R_n \triangleq \hat{\psi}_n - \psi_0$ and estimating its distribution via bootstrap gives us

Pivotal confidence interval

$$C_n = (2\hat{\psi}_n - \hat{\psi}^*_{n,1-\alpha/2}, 2\hat{\psi}_n - \hat{\psi}^*_{n,\alpha/2}) \tag{15}$$

Studentized confidence interval

We use *studentized intervals*

1. (Typically) requires nested bootstrapping for estimating $\hat{se}_b^*$

Let

$$Z^*_{n,b} = \frac{\hat{\psi}^*_{n,b} - \hat{\psi}_n}{\hat{se}_b^*} \tag{16}$$

Studentized confidence interval

$$C_n = (\hat{\psi}_n - z^*_{1-\alpha/2}\hat{se}_b, \hat{\psi}_n - z^*_{\alpha/2}\hat{se}_b) \tag{17}$$

Estimating bias

For biased estimators, we may wish to "*correct*" the bias.

- Bootstrapping allows us to estimate the bias

We can estimate the bias via

$$\hat{b} = \hat{\psi}_n - \frac{1}{B}\sum_{b=1}^{B} \hat{\psi}^*_{n,b} \tag{18}$$

And update our estimator as

$$\tilde{\psi}_n = \hat{\psi}_n + \hat{b} \tag{19}$$

Bagging

*B*ootstrap *Agg*regation

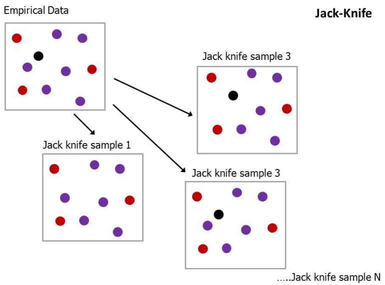
- Create B replicates of data using bootstrap
- Apply a learning method to each replicate resulting in B fits, i.e. $\hat{f}_n^{(1)}, \ldots, \hat{f}_n^{(B)}$
- Average the predictions across $\hat{f}_n^{(b)}$, i.e.

$$\hat{f}_n^{bag}(x) = \frac{1}{B} \sum_{b=1}^{B} \hat{f}_n^{(b)}(x) \tag{20}$$

Can greatly reduce the variance in estimators

- Particularly ones known for overfitting

The jackknife



A resampling method (like the Bootstrap), but

- The Bootstrap resamples data from P_n and calculates $\hat{\Psi}(\hat{P}_n^*)$
- The Jackknife leaves out (random) partitions from P_n and calculates $\hat{\Psi}(\hat{P}_n^*)$

Both methods use simulated distributions to calculate SE

The jackknife

The general algorithm (applied to our investment example):

- Assume that P_n is a good approximation of P_0 and choose a number of observations d to delete
  - where $0 < d < n$
- Iteratively:
  - Exclude d observations from our data (resulting in $P_n^{*,d}$)
  - Calculate $\hat{\alpha}_n$ from $P_n^{*,d}$ (call this $\hat{\alpha}_n^{*,d}$)
- Calculate the SD of the $\hat{\alpha}_n^{*,d}$ estimates

If $d > 1$:

$$\widehat{SE}_B(\hat{\alpha}_n) = \sqrt{\frac{n-d}{d\binom{n}{d}} \sum_z \left(\hat{\alpha}_n^{*,z} - \frac{1}{\binom{n}{d}} \sum_{z'} \hat{\alpha}_n^{*,z'} \right)^2} \tag{21}$$

When $d = 1$, this simplifies to:

$$\widehat{SE}_B(\hat{\alpha}_n) = \sqrt{\frac{n-1}{n} \sum_{i=1}^{n} \left(\hat{\alpha}_n^{*,i} - \frac{1}{n} \sum_{i'=1}^{n} \hat{\alpha}_n^{*,i'} \right)^2} \tag{22}$$

Jackknife vs Bootstrap

Some similarities:

- The Jackknife and Bootstrap are asymptotically equivalent
- The theoretical arguments proving the validity of both methods rely on large samples

Jackknife vs Bootstrap

Some similarities:

- The Jackknife and Bootstrap are asymptotically equivalent
- The theoretical arguments proving the validity of both methods rely on large samples

Some differences:

- The jackknife is less computationally expensive
- The jackknife is a linear approximation to the bootstrap
- The jackknife doesn't work well for sample quantiles like the median
- The bootstrap procedure has lots of variations
  - e.g. You can bootstrap the bootstrapped samples to try and get second-order accuracy (aka bootstrap-t)

References

[1] ISL. Chapters 5.

[2] ESL. Chapter 7.