Lecture 12: The Bootstrap

Reading: Chapter 5

STATS 202: Data mining and analysis

October 18, 2019

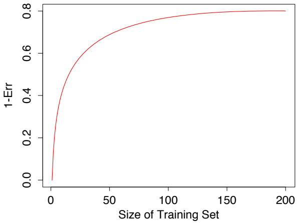
Announcements

- Midterm is on Friday, Oct 25
  - Topics: chapters 1-5 and 10 of the book — everything until and including today's lecture.
  - We will post two practice exams soon.
  - Closed book, no notes. All "hard" equations will be provided.
  - SCPD students: if you haven't chosen your proctor already, you must do it ASAP. For guidelines see:

    `http://scpd.stanford.edu/programs/courses/graduate-courses/exam-monitor-information`

The learning curve and choosing k in k-fold cross validation

The learning curve

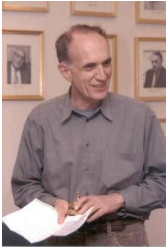


- Recall that as we increase k, we decrease the bias but increase the variance of the cross validation error.
- How does the test error change as we increase the size n of the training set? Consider the curve on the left:
  - If $n = 200$, then 5-fold CV estimates error using dataset of size $\frac{4}{5} \cdot 200 = 160$: introduces little bias!
  - If $n = 50$, then 5-fold CV estimates error using dataset of size $\frac{4}{5} \cdot 50 = 40$: introduces more bias.

Cross-validation vs. the Bootstrap

Cross-validation: principally used to estimate prediction error.

The Bootstrap: principally used to estimate various measures of error or uncertainty of parameter estimates, e.g. standard error (SE) of parameter estimates, confidence intervals for parameters.



- One of the most important techniques in all of Statistics.
- Widely applicable, extremely powerful, computer intensive method.
- Popularized by Brad Efron, from Stanford.

Standard errors in linear regression

Standard error: SD of an estimate from a sample of size n.

```
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
```

Classical way to compute Standard Errors

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

$$\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2.$$

What is the Standard Error of $\hat{\sigma}^2$?

1. Assume that $x_1, \ldots, x_n$ are i.i.d. normally distributed.
2. From that assumption one can derive that $Var(\hat{\sigma}^2) = \frac{2\sigma^4}{n-1}$, therefore $SE(\hat{\sigma}^2) = \frac{\sqrt{2}\sigma^2}{\sqrt{n-1}}$.
3. **Problem:** We typically don't know σ!
4. So assume $\frac{\hat{\sigma}^2}{\sqrt{n-1}}$ is reasonable close to $\frac{\sigma^2}{\sqrt{n-1}}$.
5. Then can use the estimate $\widehat{SE(\hat{\sigma}^2)} = \frac{\sqrt{2}\hat{\sigma}^2}{\sqrt{n-1}}$.

Limitations of the classical approach

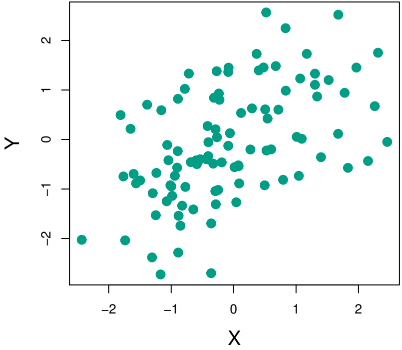
The classical approach works for certain statistics under specific modeling assumptions. However, what happens if:

- The modeling assumptions — for example, $x_1, \ldots, x_n$ being normal — break down?
- The estimator does not have a simple form and its sampling distribution cannot be derived analytically?

Example. Investing in two assets

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

These returns are observed every day: $(x_1, y_1), \ldots, (x_n, y_n)$.



Example. Investing in two assets

We have a fixed amount of money to invest and we will invest a fraction α on X and a fraction $(1-\alpha)$ on Y. Therefore, our return will be

$$\alpha X + (1-\alpha)Y.$$

Our goal will be to minimize the variance of our return as a function of α. One can show that the optimal α is:

$$\alpha = \frac{\sigma_Y^2 - \mathsf{Cov}(X,Y)}{\sigma_X^2 + \sigma_Y^2 - 2\mathsf{Cov}(X,Y)}.$$

Proposal: Use an estimate:

$$\hat{\alpha} = \frac{\hat{\sigma}_Y^2 - \hat{\mathsf{Cov}}(X,Y)}{\hat{\sigma}_X^2 + \hat{\sigma}_Y^2 - 2\hat{\mathsf{Cov}}(X,Y)}.$$

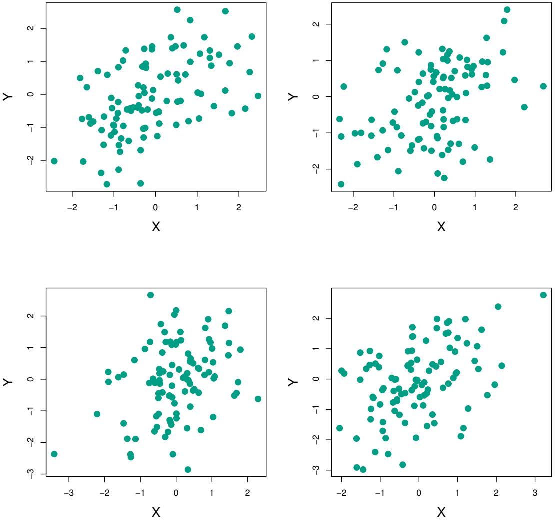
Example. Investing in two assets

Suppose we compute the estimate $\hat{\alpha} = 0.6$ using the samples $(x_1, y_1), \ldots, (x_n, y_n)$.

- How sure can we be of this value?
- If we sampled another set of observations $(x_1, y_1), \ldots, (x_n, y_n)$, would we get a wildly different $\hat{\alpha}$?

In this thought experiment, we know the actual joint distribution $P(X, Y)$, so we can resample the n observations to our hearts' content.

Resampling the data from the true distribution



Computing the standard error of $\hat{\alpha}$

Suppose we can sample as many data as we want. For each resampling of the data,

$$(x_1^{(1)}, y_1^{(1)}), \ldots, (x_n^{(1)}, y_n^{(1)})$$

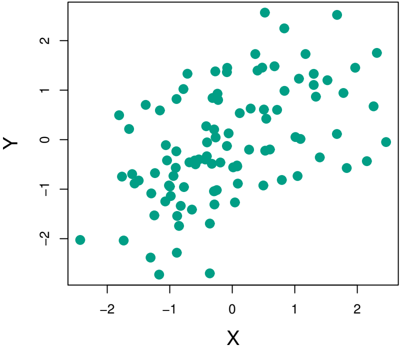
$$(x_1^{(2)}, y_1^{(2)}), \ldots, (x_n^{(2)}, y_n^{(2)})$$

$$\cdots$$

we can compute a value of the estimate $\hat{\alpha}^{(1)}, \hat{\alpha}^{(2)}, \ldots$.

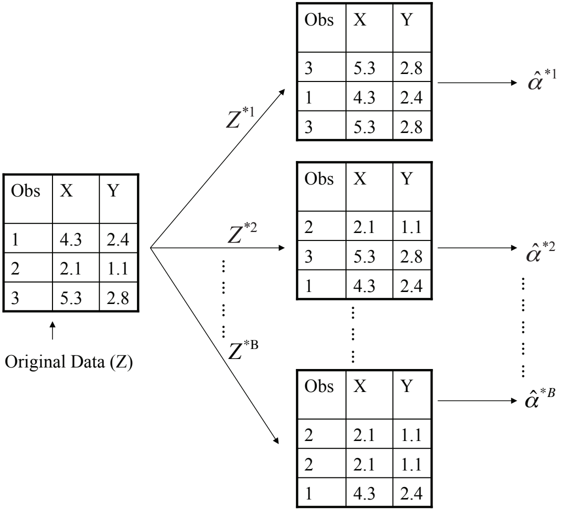
The standard deviation of these values approximates the Standard Error of $\hat{\alpha}$.

In reality, we only have one dataset of size n!



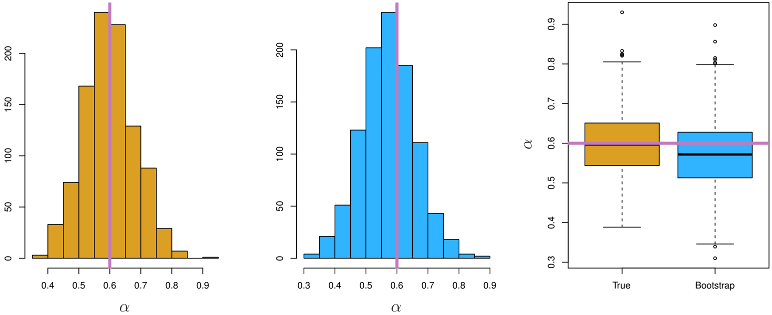
- However, this dataset can be used to approximate the joint distribution of P of X and Y by forming the *empirical distribution* $\hat{P}(X, Y)$ which gives probability $\frac{1}{n}$ to each pair (x_i, y_i).
- **The Bootstrap:** Instead of sampling new datasets from the unknown distribution P, resample from the empirical distribution $\hat{P}$.
- Equivalently, resample the data by drawing n samples *with replacement* from the actual observations.

A schematic of the Bootstrap



Each resampled dataset Z^{*b} is called a *bootstrap replicate*.

Comparing Bootstrap resamplings to resamplings from the true distribution



Bootstrapping your favorite statistics

The bootstrap is broadly applicable and can be used to estimate the SE of a wide variety of statistics including linear regression coefficients, model predictions $\hat{f}(x_0)$, principal component loadings,....