Bootstrap, Jackknife and cross-validation

October 8, 2013
Bootstrap, jackknife and cross-validation

- Computer intensive resampling methods of statistical inference (point estimation, standard error, confidence interval, model selection, hypothesis testing)
- Jackknife for eliminating bias of an estimate
- Bootstrap for estimating standard error of an estimate and for computing confidence intervals for population parameters
- Cross-validation for model selection
Quenouille (1949) introduced the method in order to eliminate bias in an estimate.

Tukey (1958) suggested that Quenouille’s method could be used as a non-parametric method of estimating the mean and variance of an estimate.

→ the term **jackknife** to signify an all-purpose statistical tool.
Jackknife algorithm for bias reduction

1. Estimate $\theta$ according to some appropriate algorithm (for example ML or LS) to obtain an estimate $\hat{\theta}$.

2. Delete an observation from the data and recalculate the estimate for $\theta$ from the remaining $n - 1$ observations. The estimate is denoted by $\hat{\theta}_{-i}$.

3. Calculate the “pseudovalue”

\[ S_i = n\hat{\theta} - (n - 1)\hat{\theta}_{-i} \]

4. Repeat steps 2 and 3 until you have gone through all observations.

5. The jackknife estimate for $\theta$ is the mean of the pseudovalues, namely

\[ \tilde{\theta} = \frac{1}{n} \sum_{i=1}^{n} S_i = n\hat{\theta} - \frac{n - 1}{n} \sum_{i=1}^{n} \hat{\theta}_{-i} \]
Conclusions

- Computationally easy
- Introduced for bias elimination but often used also for estimating standard error
- Here we introduced the so-called “delete-one jackknife”. By deleting more than one observation each time, a higher order jackknife estimators can be constructed
Bootstrap

- Resampling with replacement, introduced as a general method by Efron (1979)
- The origin of the name *bootstrap*: someone pulling himself out of mud with his bootstraps
- Mostly used to find confidence intervals for population parameters
In the absence of any other knowledge about a population, the distribution of values found in a random sample of size $n$ from the population is the best guide to the distribution in the population.

To approximate what would happen if the population was resampled, we resample the data.

The sampling is done with replacement.
Bootstrap in point estimation

1. A set of $n$ observations from which we estimate a parameter $\theta$, the estimate is denoted by $\hat{\theta}$

2. Generate a bootstrap sample: randomly sample with replacement $n$ observations from the original data set and calculate the estimate of $\theta$ denoted by $\theta^*_i$

3. Repeat 2) many times, say $B$, and obtain the estimates $\theta^*_1, \theta^*_2, ..., \theta^*_B$
Simplest bootstrap estimate for $\theta$ would be the mean of the bootstrap estimates:

$$\theta^* = \frac{1}{B} \sum_{i=1}^{B} \theta_i^*$$

Bias-adjusted bootstrap estimate: If $\hat{\theta}$ is biased for $\theta$, then $\theta^*$ will itself be biased because it estimates $\hat{\theta}$ rather than $\theta$. The bias is defined as

$$\text{Bias}(\theta) = \hat{\theta} - \theta$$

and it can be estimated by

$$\hat{\text{Bias}}(\theta) = \theta^* - \hat{\theta}$$

By combining the two formulae above we get the bias-adjusted bootstrap estimate

$$\theta_{BA}^* = \hat{\theta} - \hat{\text{Bias}}(\theta) = \hat{\theta} - \theta^* + \hat{\theta} = 2\hat{\theta} - \theta^*$$
Estimating standard error

To be able to compute the standard error of an estimate
- sampling distribution of the statistic has to be known
- the exact formula for the standard error has to be known

The standard error of the estimate $\hat{\theta}$ can always be estimated by the sample standard deviation of the $B$ bootstrap replications

$$\hat{s}_e_B = \sqrt{\frac{1}{B-1} \sum_{i=1}^{B} (\theta_i^* - \theta^*)^2},$$

where $\theta^*$ is the mean of the bootstrap estimates (or by using the bias-adjusted estimates).

**Remark:** To obtain a good estimate for the standard error of the estimate, 100 bootstrap replicates may be enough, but in order to get good confidence intervals, a much larger number of replicates (say 1000) is needed.
Example: Study of snow geese

- **Aim:** To study feeding habits of baby snow geese

- **Data collection:** 33 goslings were without food until their guts were empty and then, they were allowed to feed for 6 hours on a diet of plants. The change in the weight of the gosling after 2.5 hours was recorded as a percentage of initial weight. Digestion efficiency (measured as a percentage) was also recorded.

- **Question:** Are the weight change and digestion efficiency correlated?
Example: Data

Baby snow geese study

Bootstrap, Jackknife and cross-validation
Example: correlation coefficient

- The computed correlation coefficient is $\hat{r} = 0.309$. How accurate is this estimate?
- Compute the standard error (s.e.) of the correlation coefficient. If it is small, the estimate is accurate.
- There is no formula for the s.e. of the correlation coefficient and therefore, it is hard to assess the accuracy of the estimate → Bootstrap (or jackknife)
We draw a bootstrap sample, a random sample of size $n$ (here 33), with replacement from the observed pairs of values, and calculate the correlation coefficient from this sample, say $r_1^*$. This is repeated a number of times, here $B = 1000$. The bootstrap estimate of the standard error of the correlation coefficient is then

$$se(\hat{r}) = \sqrt{\frac{1}{999} \sum_{i=1}^{1000} (r_i^* - r^*)^2} = 0.191,$$

where $r^*$ is the mean of the bootstrap estimates of the correlation coefficient.
Several methods exist to construct bootstrap confidence intervals.

Different methods do not necessarily give the same interval, particularly if the distribution is highly skewed.

It is difficult to know beforehand which method to use and one may have to perform a simulation study in order to choose a suitable method.
Method 1: Standard bootstrap interval

If we can assume that the estimator is (approximately) normally distributed, i.e. $\hat{\theta} \sim N(\theta, \sigma)$, a $100(1 - \alpha)\%$ confidence interval for $\theta$ is the “usual”, i.e.

$$
\hat{\theta} - z_{\alpha/2}\sigma < \theta < \hat{\theta} + z_{\alpha/2}\sigma,
$$

where $z_{\alpha/2}$ is the point for which $\text{Prob}(Z \geq z_{\alpha/2}) = \alpha/2$ for $Z \sim N(0, 1)$. 

With the standard bootstrap confidence interval $\sigma$ is estimated by the standard deviation of the estimates of a parameter $\theta$ that are found by bootstrap resampling of the values in the original sample of data. The interval is then

$$\text{Estimate} \pm z_{\alpha/2}(\text{Bootstrap standard deviation}).$$

For example, using $z_{\alpha/2} = 1.96$ gives the standard 95% bootstrap interval
It is appropriate to calculate the standard bootstrap interval if
(a) $\hat{\theta}$ is approximately normally distributed
(b) $\hat{\theta}$ is unbiased (its mean value for repeated samples from the
population of interest is $\theta$)
(c) bootstrap resampling gives a good approximation to $\sigma$

Under the conditions above the method is potentially useful
whenever an alternative method for approximating $\sigma$ is not easily
available.

In practice it may be possible to avoid requirement $b)$ by
estimating bias in $\hat{\theta}$ as part of the bootstrap procedure.
A 95% bootstrap confidence interval for the correlation coefficient is

\[ \hat{\theta} \pm z_{\alpha/2} \hat{\text{se}}_B = 0.309 \pm 1.96 \cdot 0.191 = 0.309 \pm 0.374 \]

giving us the interval

\[ [-0.065, 0.683]. \]

According to this result there is no significant correlation between the weight chance and digestion efficiency of snow geese (the confidence interval contains 0).
Example: Bootstrap distribution of $\hat{r}$
The distribution of bootstrap replicates is a description of the
distribution of the parameter of interest.
Rather than using the normality of the estimator, we can use
the distribution itself to assign the lower and upper confidence
limits.
The bootstrap replicates are ranked from lowest to highest
(assuming no ties) and the lower and upper limits are chosen
as follows

- **Lower limit**: choose the bootstrap replicate at which $100 \frac{\alpha}{2} \%$
of the replicates lie below that value
- **Upper limit**: choose the bootstrap replicate at which $100 \frac{\alpha}{2} \%$
of the replicates lie above that value

If the bootstrap distribution is symmetric then the limits will
be symmetric about the estimate, otherwise asymmetric
Example: Percentile intervals

The new confidence interval for the correlation coefficient is $[-0.042, 0.589]$
Other bootstrap confidence intervals

- Standard bootstrap confidence intervals have good theoretical coverage probabilities but tend to be erratic in actual practice.
- Percentile intervals are less erratic, but have less satisfactory coverage properties.
- Other possibilities
  - Bias-corrected percentile method (corrects for the initial sample being a biased sample of the distribution)
  - Accelerated and bias-corrected percentile method (as the bias-corrected method but allows the standard error of $\hat{\theta}$ vary)
Bootstrap is mostly used for estimating standard error of an estimate or for finding confidence intervals for parameters.

Can also be used in hypothesis testing.

Jackknife can be thought as an approximation of bootstrap but it is less effective than bootstrap (only $n$ samples).

There is also a parametric version of bootstrap (fit a parametric model to the data and use this model to produce bootstrap samples).
Cross-validation

- Originally suggested by Kurtz (1948)
- Extended to double cross-validation by Mosier (1959)
- Original objective was to verify replicability of results, i.e. to find out whether a result is replicable or just obtained by a chance
Prediction error

- Measures how well a model predicts the response value of a future observation
- Used for model selection
- In regression models prediction error is defined as the expected square difference between a future response and the prediction of the model in regression models. i.e.

\[ \mathbb{E}(y - \hat{y})^2 \]
Linear regression model

- We have a set of points \( \{(x_i, y_i)\}, \ i = 1, \ldots, n \)
- We are interested in the unknown relation between \( y_i \) and \( x_i \), e.g.
  \[
y_i = \sum_{j=0}^{p} \beta_j x_i^j + \epsilon_i
  \]
- Error terms \( \epsilon_i \) often assumed to be independent and \( N(0, \sigma^2) \) distributed
Model selection problem

We fit a first order linear model (left) and a quadratic model (right) to the same data \((n=10)\)
Which model is the best?

- I.e. how well are we going to predict future data drawn from the same distribution?
- As a prediction error, we can compute the average

\[
\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2
\]

where \( \hat{y}_i \) is the prediction by the model (either first order or quadratic)

- Here, \( y_i \) taken from the sample in hand, and is not a future response.
- What is the prediction error for a new sample? How to get a new sample to compute this prediction?
Cross validation

- To obtain a more realistic estimate of the prediction error, we need a test sample that is different from the training sample.
- One can, for example, split the observations \( \{(x_i, y_i)\}, \ i = 1, \ldots, n \), into two sets, one for training and one for testing → cross-validation
Cross validation

- **Simple cross-validation**: Divide the data into two groups, one for training and one for testing. The parameters $\beta$ are estimated from the training set. The cross-validation is the prediction error computed using the test set.

- **Double cross-validation**: Models are estimated for both sub-samples, and then both equations are used to generate cross-validation.
K-fold cross validation

- Data divided into \( K \) subsets (depends on \( n \) into how many)
- For each subset \( i \) (repeat \( K \) times)
  - Estimate the model based on the remaining \( K - 1 \) subsets (without the subset \( i \))
  - Compute the prediction error for the subset \( i \)
- Compute the average of the \( K \) prediction errors as the overall estimate of the prediction error
- \( K \) large: prediction error can be estimated accurately but the variance of the estimator may be large.
- For large data sets, even 3-fold cross-validation is accurate. For small data, we may have to use leave-one-out cross-validation where \( K = n \)
Conclusions

- Used in model selection in regression
- Can be used also e.g. in classification ($y_i$; a label indicating class and the prediction error is defined as a missclassification rate)


