Monte Carlo methods for Statistical Inference: Resampling

Hung Chen

hchen@math.ntu.edu.tw

Department of Mathematics

National Taiwan University

17th March 2004

Meet at NS 104 On Wednesday from 9:10 to 12.

Outline

- Introduction
  - Nonparametric Bootstrap
  - Classical Paradigm on Inference
  - Error in Bootstrap Inference
  - R-programming
- Applications of Bootstrap
  - Confidence Intervals
  - Regressions
  - Hypothesis Tests
  - Censored data
- Resampling Methods

  - **Permutation Tests**
  - **The Jackknife**
  - **Cross Validation and Model Selection**

- **References**
  - **Bickel, P. and Freedman, D. (1981) Some asymptotic theory for the bootstrap.** *Annals of Statistics*, 9, **1196-1217.**
  - **Davison, A.C. and Hinkley, D.V. (1997).** *Bootstrap Methods and their Application*. **Cambridge: Cambridge University Press.**
  - **DiCicco, T.J and Efron, B. (1996). Bootstrap confidence intervals.** *Statistical Science*, 11, **189-228.**
  - **Efron, B. (1979) Bootstrap Methods: Another**

Look at the Jackknife. *Annals of Statistics* **7 1¡X26.**

- **Efron, B. and R.J. Tibshirani (1993)** *An Introduction to the Bootstrap*. **New York: Chapman and Hall.**
- **http://www-stat.stanford.edu/∼susan/courses/s208/**
- **Refer to Rnews 2002 December issue on** *Resampling Methods in R: The boot Package*.

Bootstrapping

Bootstrapping is a general approach to statistical inference based on building a sampling distribution for a statistic by resampling from the data at hand.

- The term bootstrapping,due to Efron (AS, 1979), is an allusion to the expression pulling oneself up by one's bootstraps.
- It uses the sample data as a population from which repeated samples are drawn.
- Two R libraries for bootstrapping are associated with extensive treatments of the subject:
  - Efron and Tibshirani's (1993) bootstrap library
  - Davison and Hinkley's (1997) boot library

- **There are several forms of the bootstrap, and, additionally, several other resampling methods that are related to it, such as** *jackknifing*, *cross-validation*, *randomization tests*, **and** *permutation tests*.

Nonparametric Bootstrapping

Scientific question:

- Why Sample?
  - Sampling can provide reliable information at far less cost than a census.
  - Samples can be collected more quickly than censuses.
  - Sampling can lead to more accurate estimates than censuses (!).
    - When samples are used a small number of well-trained interviewers can spend more time getting high quality responses from a few sampled people.

- With probability sampling, you can quantify the sampling error from a survey.
- Sampling is necessary when a unit must be destroyed to be measured (e.g., breaking apart a Chips Ahoy! Cookie to measure the number of chocolate chips)

- Suppose that we draw a sample $S = \{X_1, X_2, \ldots, X_n\}$ from a population $\mathrm{P} = \{x_1, x_2, \ldots, x_N\}$ where N is very much larger than n.
- Abstraction: Suppose that with any design, with or without replacement, the probability of including unit i in the sample is π_i (> 0), for $i = 1, 2, \ldots, N$.
  - The Horvitz-Thompson (1952) estimator for the

population total X_T is defined as

$$\hat{Y}_{HT} = \sum_{i \in S} \frac{x_i}{\pi_i}$$

where S contains the distinct units only and hence the size of S could be less than the number n of units drawn.

- If the sampling is without replacement, the size of S must be n.
- Under a sampling with replacement, it can be shown that for a fixed sample size n,

$$\pi_i = 1 - (1 - p_i)^n$$

where p_i is the probability of selecting the ith unit of the population on each draw.

  - Under a simple random sampling without replacement, it can be shown that for a fixed sample size n, $\pi_i = n/N$.

- For simplicity, think of the elements of the population as scalar values, but they could just as easily be vectors (i.e., multivariate).

- Now suppose that we are interested in some statistic $T = t(S)$ as an estimate of the corresponding population parameter $\theta = t(\mathbf{P})$.
  - θ could be a vector of parameters and T the corresponding vector of estimates.
  - In inference, we are interested in describing the random variable $t(S) - t(\mathbf{P})$ which varies with S.
    - Find the distribution of $t(S) - t(\mathbf{P})$.

  - **How do we describe the distribution of $t(S) - t(\mathbf{P})$?**
  - **Chebyschev's inequality, Asymptotic analysis, ...**

Essential idea of the nonparametric bootstrap is as follows:

- **Draw a sample of size n from among the elements of S, sampling with replacement.**
  Call the resulting bootstrap sample $S_1^* = \{X_{11}^*, X_{12}^*, \ldots, X_{1n}^*$
- **In effect, we are treating the sample S as an estimate of the population P; that is, each element X_i of S is selected for the bootstrap sample with probability $1/n$, mimicking the original selection of the sample S from the population P.**
- **Repeat this procedure a large number of times, P,**

selecting many bootstrap samples; the bth such bootstrap sample is denoted $S_b^* = \{X_{b1}^*, X_{b2}^*, \ldots, X_{bn}^*\}$. The population is to the sample as the sample is to the bootstrap samples.

- Compute the statistic T for each of the bootstrap samples; that is $T_b^* = t(S_b^*)$.
  - Let B denote the number of times on resampling.
  - In theory, we can determine the distribution of $T^* - T$ when $B \to \infty$.
  - We are doing simulation essentially.
- Suppose that we observe the sample

$$\mathbf{X} = X_1, \ldots, X_n \overset{iid}{\to} F(\theta),$$

indexed by unknown parameter θ and compute the

statistic

$$\hat{\theta} = \theta(X_1, \ldots, X_n) = \theta(\mathbf{X}).$$

- **Denote the empirical distribution by**

$$F_n(x) = \frac{\#\{X_i \leq x\}}{n}.$$

- **Think of the parameter $\theta = \theta(F)$ and the statistic $\hat{\theta}$ as functionals $\hat{\theta}(F_n)$.**
- **The idea of bootstrap is to exploit the analogy**

$$\theta(F_n) - \theta(F) : \theta(F_n^*) - \theta(F_n)$$

where F_n^* denotes the empirical distribution of a sample from F_n.

- **How do we find the sampling distribution of $\bar{X} - E(X)$ when $X_1, \ldots, X_n$ are from exponential distribution?**

  - How do I utilize the information available to me?
  - Think of parametric bootstrap.

- In the parametric bootstrap, the distribution function of the population of interest, F, is assumed known up to a finite set of unknown parameters θ.
  - $\hat{F}$ is F with θ replaced by its sample estimate (of some kind).

- Algorithm of parametric bootstrap:
  - Obtain estimates of the parameters that characterize the distribution within the assumed family.
  - Generate B random samples each of size n from the estimated distribution, and for each sample, compute an estimator T_b^* of the same functional form as the original estimator T.

- The distribution of the T_b^*'s is used to make inferences about the distribution of T.

- Assume that the distribution of T_b^* around the original estimate T is analogous to the sampling distribution of the estimator T around the population parameter θ.
  - Consider the problem of correcting the bias of T as an estimate of θ.
    - Let B denote the number of times of resampling.
    - The average of the bootstrapped statistics is

$$\bar{T}^* = \hat{E}^*(T^*) = \frac{\sum_{b=1}^{B} T_b^*}{R}.$$

    - Bootstrap estimate of the bias would be $\bar{T}^* - T$.

    - $*$ Recall that the bias is $E(T) - \theta$.
    - $*$ Improve the estimate T by $T - (\bar{T}^* - T)$.
  - – How do we estimates the sampling variance or sampling distribution of T?

Note that the random selection of bootstrap samples is not an essential aspect of the nonparametric bootstrap:

- At least in principle, we could enumerate all bootstrap samples of size n. Then we could calculate $E^*(T^*)$ and $Var^*(T^*)$ exactly, rather than having to estimate them.
- The number of bootstrap samples n^n, however, is astronomically large unless n is tiny.

Error in Bootstrap Inference

There are two sources of error in bootstrap inference:

(1) the error induced by using a particular sample S to represent the population

(2) the sampling error produced by failing to enumerate all bootstrap samples.

- **This source of error can be controlled by making the number of bootstrap replications R sufficiently large.**

- **A traditional approach to statistical inference is to make assumptions about the structure of the population (e.g., an assumption of normality), and, along with the stipulation of random sampling, to use**

these assumptions to derive the sampling distribution of T, on which classical inference is based.

- **In certain instances, the exact distribution of T may be intractable, and so we instead derive its asymptotic distribution.**
- **This familiar approach has three potentially important deficiencies:**
  1. **If the assumptions about the population are wrong, then the corresponding sampling distribution of the statistic may be seriously inaccurate.**
  2. **If asymptotic results are relied upon, these may not hold to the required level of accuracy in a relatively small sample.**
  3. **The approach requires sufficient mathematical**

prowess to derive the sampling distribution of the statistic of interest. In some cases, such a derivation may be prohibitively difficult.

- Background:
  Some of the theory involves functional Taylor expansions.
  $$\hat{\theta} - \theta \approx \theta^{'}(F)(F_n - F),$$
  where $F_n - F$ can be approximated by a Brownian bridge.
  By the same reasoning,
  $$\hat{\theta}^{*} - \hat{\theta} \approx \theta^{'}(F_n)(F_n^{*} - F_n).$$
  Again, $F_n^{*} - F_n$ can be approximated by a Brownian bridge.
- If this statistics is sufficiently smooth so that the

functional derivatives $\theta^{'}(F) \approx \theta^{'}(F_n)$, then the procedure gets the right SE.

R-programming

In this demonstration, we consider estimating the parameter θ of an exponential distribution.

- **Data generation with $\theta = 2$:**

```
options(digits=3) & x<- rexp(20,2)
print(theta<- sd(x))
```

- **Parametric bootstrap**

```
lambda<- 1/mean(x) & thetas<- 1:1000
for (i in 1:1000)
  thetas[i]<- sd(rexp(30,lambda))
c(mean(thetas),sd(thetas))
quantile(thetas,c(0.025,0.975))
theta-rev(quantile(thetas-theta,c(0.025,0.975)))
```

- **Nonparametric bootstrap**

```
thetas2<- 1:1000
for (i in 1:1000)
  thetas[i]<- sd(sample(x,repl=T))
c(mean(thetas2),sd(thetas2))
quantile(thetas2,c(0.025,0.975))
theta-rev(quantile(thetas2-theta,c(0.025,0.975)))
```

- **Monte Carlo estimate (knowing the truth)**

```
thetas3<- 1:10000 for (i in 1:100000)
  thetas3<- sd(rexp(30,2))
c(mean(thetas3),sd(thetas3))
```

Bootstrap Confidence Intervals

There are several approaches to constructing bootstrap confidence intervals.

- **The normal-theory interval assumes that the statistic T is normally distributed (which is often approximately the case for statistics in sufficiently large samples), and uses the** *bootstrap estimate of sampling variance*, **and perhaps of** *bias*, **to construct a $100(1-\alpha)$-percent confidence interval of the form**

$$\theta = (T - \hat{B}^*) \pm z_{1-\alpha} \hat{SE}^*(T^*).$$

  Here,

  - $\hat{B}^*$ **and** $\hat{SE}$ **are the bootstrap estimate of the bias and standard error of** T**.**

- $z_{1-\alpha/2}$ **is the $1-\alpha/2$ quantile of the standard-normal distribution.**

- **Bootstrap percentile interval: It is to use the empirical quantiles of T_b^* to form a confidence interval for θ.**

$$T^*_{(lower)} < \theta < T^*_{(upper)}$$

where $T^*_{(1)}, T^*_{(2)}, \ldots, T^*_{(B)}$ are the ordered bootstrap replicates of the statistic; lower = $[(B+1)\alpha/2]$; upper = $[(B+1)(1-\alpha/2)]$; and the square brackets indicate rounding to the nearest integer.

  - **Although they do not artificially assume normality, percentile confidence intervals often do not perform well.**
    - **Do a transformation!**

- $*$ **We have a lot of experience with approximate transformations to normality.**
- $*$ **Suppose there is a monotonically increasing transformation g and a constant τ such that the random variable**

  $$Z = c[g(\hat{\theta}^*) - g(\theta)]$$

  has a symmetric distribution about zero.
- $*$ **Let H be the distribution function of Z. Then**

  $$G_{\hat{\theta}^*}(t) = H(c[g(t) - g(\theta)]),$$

  and

  $$\hat{\theta}^{*(1-\alpha)} = g^{-1}(g(\hat{\theta}^*) + z^{(1-\alpha)}/c),$$

  where $z^{(1-\alpha)}$ is the $1-\alpha$ quantile of Z.

– **How do we employ this concept on vector-valued parameters?**

  - $*$ Do we restrict to an ellipsoidal region?
  - $*$ If it is, the shape is determined by the covariances of the estimators.
  - $*$ How about the one-sided confidence interval?

- **The bootstrap t interval: Determine the following distribution by the bootstrap method.**

$$\frac{\hat{\theta}^* - \hat{\theta}}{s(\hat{\theta}^*)}$$

- **Bias-corrected, accelerated (or BC_a) percentile intervals:**
  Steps:

– **Calculate**

$$z = \Phi^{-1}\left[\frac{\sum_{b=1}^{B} 1(T_b^* \leq T)}{B+1}\right]$$

where $\Phi^{-1}(\cdot)$ is the standard-normal quantile function.

– **If the bootstrap sampling distribution is symmetric, and if T is unbiased, then this proportion will be close to 0.5, and the correction factor z will be close to 0.**

– **Let $T_{(-i)}$ represent the value of T produced when the ith observation is deleted from the sample; there are n of these quantities.**
Let $\bar{T}$ represent the average of the $T_{(-i)}$.

Calculate

$$a = \frac{\sum_{i=1}^{n}[T_{(-i)} - \bar{T}]^3}{6\left[\sum_{i=1}^{n}(T_{(-i)} - \bar{T})^2\right]^{3/2}}.$$

– **With the correction factors z and a in hand, compute**

$$\begin{aligned} a_1 &= \Phi\left[z + \frac{z - z_{1-\alpha/2}}{1 - a(z - z_{1-\alpha/2})}\right] \\ a_2 &= \Phi\left[z + \frac{z + z_{1-\alpha/2}}{1 - a(z - z_{1-\alpha/2})}\right]. \end{aligned}$$

– **The values a_1 and a_2 are used to locate the endpoints of the corrected percentile confidence in-**

terval:

$$T_{(lower^*)} < \theta < T_{(upper^*)}$$

where $lower^* = [Ba_1]$ **and** $upper^* = [Ba_2]$**.**
When the correction factors a **and** z **are both** 0**, it corresponds to the (uncorrected) percentile interval.**

Bootstrapping Regressions

Consider $(x_1, x_2, y)_i$ **for** $i = 1, \ldots, n$**.**

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

with $E(\epsilon) = 0$ **and** $Var(\epsilon) = \sigma^2$**.**
Assume the ϵ_i**s are independent.**

- **Parametric bootstrap:**
  - **Obtain** $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$
  - **Sample** $\epsilon_i^* \sim$ **iid N(0,**σ^2**)**
  - **Take** $y_i^* = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \epsilon_i^*$
  - **Obtain** $\hat{\beta}_0^*$, $\hat{\beta}_1^*$, $\hat{\beta}_2^*$ **and repeat many times**
- **Resampling residuals:**
  - **Obtain** $\hat{\beta}_0$, $\hat{\beta}_1$, **and** $\hat{\beta}_2$**.**

- **Calculate residuals** $\epsilon_i^* = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2}$
- **Sample** ϵ_i^* **by drawing with replacement from** $\{\hat{\epsilon}_i\}$
- **Add** ϵ_i^* **to** $\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2}$**.**
- **Obtain** $\hat{\beta}_0^*$**,** $\hat{\beta}_1^*$**,** $\hat{\beta}_2^*$ **and repeat many times.**
- **It assumes that the distribution of** ϵ **is the same in all regions of the model.**
- **Better efficiency but is not robust to getting the wrong model.**

- **Resampling cases:**
  - **Sample** $(x_{i1}^*, x_{i2}^*, y_i^*)_i$ **by drawing with replacement from** $(x_{i1}, x_{i2}, y)_i$**.**
  - **Obtain** $\hat{\beta}_0^*$**,** $\hat{\beta}_1^*$**,** $\hat{\beta}_2^*$ **and repeat many times.**
  - **It is less efficient but it preserves the relationship**

between Y and (X_1, X_2) is better. (Think of the case that the variance is not homogeneous.)

Resampling Censored Data

In many practical settings, data is censored and so the usual bootstrap is not applicable.

- **Random right-censored data:**
  Such data is typically comprised of the bivariate observations (Y_i, D_i) where
  $$Y_i = \min(X_i, C_i) \quad D_i = I(X_i \le C_i)$$
  where $X_i \sim F$ and $C_i \sim G$ independently and $I(A)$ is the indicator function of the event A.
  - **Example: remission times for patients with a type of leukaemia**
    - **The patients were divided into those who received maintenance chemotherapy and those who did not.**

- We are interested in the median remission time for the two groups.

- survfit: Computes an estimate of a survival curve for censored data using either the Kaplan-Meier or the Fleming-Harrington method or computes the predicted survivor function for a Cox proportional hazards model.

-
```
data(aml, package="boot")
fit<- survfit(Surv(time,group)~cens,data=aml)
plot(fit)
```

- **Algorithm 1:**

  - Nonparametric estimates of F and G are given by the Kaplan-Meier estimates $\hat{F}$ and $\hat{G}$, the latter being obtained by replacing d_i by $1 - d_i$.

- $\theta = F_T^{-1}(0.5) - F_C^{-1}(0.5)$
- **Sampling $X_1^*, \ldots, X_n^*$ from $\hat{F}$ and independently sampling $C_1^*, \ldots, C_n^*$ from $\hat{G}$.**
- **(Y_i^*, D_i^*) can then be found from (X_i^*, C_i^*) in the same way as for the original data.**
- **Efron (1981) showed that this is identical to resampling with replacement from the original pairs.**

```
aml.fun<- function(data){
 surv<-survfit(Surv(time,group)~cens,data=data)
 out<- NULL
 st <- 1
 for (s in 1:length(surv$strata)) {
 inds <- st:(st+surv$strata[s]-1)
 md<-min(surv$time[inds[1-surv$surv[inds]>=0.5]]
  st <- st+surv$strata[s]
  out<- c(out,md)
     }
  }
aml.case<- censboot(aml,aml.fun,R=499,strata
          =aml$group)
```

  - Refer to the R function censboot.

- Conditional bootstrap:
  - This approach conditions the resampling on the observed censoring pattern since this is, in effect, an ancillary statistic.
  - Sample $X_1^*, \ldots, X_n^*$ from $\hat{F}$ as before.
  - If the ith observation is censored then we set

$C_i = y_i$ and if it is not censored we sample an observation from the estimated conditional distribution of C_i given that $C_i > y_i$.

- **Having thus obtained $X_1^*, \ldots, X_n^*$ and $C_1^*, \ldots, C_n^*$ we proceed as before.**
- **Technical problem: Suppose that the maximum value of $y_1, \ldots, y_n$, y_k say, is a censored observation. Then $X_k^* < y_k$ and $C_k^* = y_k$ so the bootstrap observation will always be uncensored.**

  Alternatively, if the the maximum value is uncensored, the estimated conditional distribution does not exist.

  In order to overcome these problems we add one extra point to the data set which has an observed value greater than $\max\{y_1, \ldots, y_n\}$ and has the op-

posite value of the censoring indicator to the maximum.

– **R-program**

```
aml.s1<-survfit(Surv(time,cens)~group,data=aml)
aml.s2<-survfit(Surv(time-0.001*cens,1-cens)~1,
     data=aml)
aml.cond<-censboot(aml,aml.fun,R=499,strata=
   aml$group,F.surv=aml.s1,G.surv=aml.s2,
   sim="cond")
```

Bootstrap Hypothesis Tests

It is also possible to use the bootstrap to construct an empirical sampling distribution for a test statistic.

- **To be added later on.**

Permutation Tests

Randomization Methods:

When an hypothesis of interest does not have an obvious test statistic, a randomization test may be useful.

- **Compare an observed configuration of outcomes with all possible configurations.**
- **The randomization procedure does not depend on assumptions about the data generating process, so it is usable in a wide range of applications.**
- **In most situations the null hypothesis for the test is that all outcomes are equally likely, and the null hypothesis is rejected if the observed outcome belongs to a subset that has a low probability under the null hypothesis, but a relatively higher probability under**

the alternative hypothesis.

- Suppose that we want to test whether the means of two data generating processes are equal.
  - The decision would be based on observations of two samples of results using the two treatments.
  - There are several statistical tests for this null hypothesis, both parametric and nonparametric, that might be used.
  - Most tests would use either the differences in the means of the samples, the numbers of observations in each sample that are greater than the overall mean or median, or the overall ranks of the observations in one sample.
  - Any of these test statistics could be used as a test

statistic in a randomization test.

Test the difference in the two sample means

Consider two samples, $x_1, x_2, \ldots, x_m$ and $y_1, y_2, \ldots, y_n$. The chosen test statistic is $t_0 = \bar{x} - \bar{y}$.

- Without making any assumptions about the distributions of the two populations, the significance of the test statistic (that is, a measure of how extreme the observed difference is) can be estimated by considering all configurations of the observations among the two treatment groups.
  - This is done by computing the same test statistic for each possible arrangement of the observations, and then ranking the observed value of the test statistic within the set of all computed values.
  - Consider a different configuration of the same set

of observations, $y_1, x_2, \ldots, x_m$ and $x_1, y_2, \ldots, y_n$ in which an observation from each set has been interchanged with one from the other set.

The same kind of test statistic, namely the difference in the sample means, is computed. Let t_1 be the value of the test statistic for this combination.

- **Consider all possible different configurations, in which other values of the original samples have been switched.**

  Compute the test statistic. Continuing this way through the full set of x's, we would eventually obtain $C(n+m, n)$ different configurations, and a value of the test statistic for each one of these artificial samples.

- **Consider the set of computed values to be a real-**

ization of a random sample from that distribution under the null hypothesis.

- The empirical significance of the value corresponding to the observed configuration could then be computed simply as the rank of the observed value in the set of all values.
- Because there may be a very large number of all possible configurations, we may wish to sample randomly from the possible configurations rather than considering them all.
- When a sample of the configurations is used, the test is sometimes called an *approximate randomization test*.

- Randomization tests have been used on small data sets for a long time. Refer to Fisher's famous *lady tasting tea* experiment in which a randomization test is used.

- Refer to Fisher (1935). Because such tests can require very extensive computations, their use has been limited until recently.
- Fisher's randomization test: Fisher (1935) gave a permutation justification for the usual test for n paired observations.
  - In his example (Darwin's Zea data) y_i and $d_i =| x_i - y_i |$ were real numbers representing plant height for treated and untreated plants.
  - Darwin conducted an experiment to examine the superiority of cross-fertilized plants over self-fertilized

plants.

  - 15 pairs of plants were used. Each pair consisted of one cross-fertilized plant and one self-fertilized plant which germinated at the same time and grew in the same pot.
  - The plants were measured at a fixed time after planting and the difference in heights between the cross- and self-fertilized plants are recorded in eighths of an inch.
  - This data can be found in the package of boot with name $darwin$.

- Darwin had calculated the mean difference.
- Fisher gave a way of calibrating this by calculating

$$S_n = \epsilon_1 d_1 + \cdots + \epsilon_n d_n$$

and considering all 2^{n+1} possible sums $\epsilon = \pm 1$ with S_0.

- **Manly, B.F.J. (1997)** *Randomization, bootstrap and Monte Carlo method in biology*, **2nd ed. Chapman & Hall, London.**

The Jackknife

Jackknife methods make use of systematic partitions of a data set to estimate properties of an estimator computed from the full sample.

- **Quenouille (1949, 1956) suggested the technique to estimate (and, hence, reduce) the bias of an estimator $\hat{\theta}_n$.**
- **Tukey coined the term** *jackknife* **to refer to the method, and also showed that the method is useful in estimating the variance of an estimator.**
- **Suppose, we have a random sample, $X_1, X_2, \ldots, X_n$, from which we compute a statistic T as an estimate of a parameter θ in the population from which the sample was drawn.**

In the jackknife method,

- **Partition the given data set into r groups each of size k. (For simplicity, we will assume that the number of observations n is kr.)**
- **Remove the jth group from the sample, and compute the estimate, T_{-j} from the reduced sample.**
- **Consider $T_j^* = rT - (r-1)T_{-j}$ which is called** *pseudovalues*. **The mean of the pseudo values, $J(T)$, is called the** *jackknife estimator* **corresponding to T:**

$$J(T) = rT - (r-1)\frac{\sum_{j=1}^{r} T_{-j}}{r}.$$

- **In most applications, it is common to take $k=1$ or $r=n$.**

- **$J(T)$ can provide information about the variance of**

the bias of the estimator T.

- **The Jackknife Bias Correction:**
  - **Suppose that we can express the bias of $\hat{\theta}_n$ as a power series in n^{-1}.**

    $$\hat{\theta}_n - \theta = \frac{a_1}{n} + \frac{a_2}{n^2} + \frac{a_3}{n^3} + \cdots$$

    where the numerators are unknowns depending on the real distribution F.

– **For $J(T)$, we have**

$$\sum_{q=1}^{\infty} \frac{a_q}{n^q} + (n-1)\left(\sum_{q=1}^{\infty} \frac{a_q}{n^q} + \theta\right)$$

$$-(n-1)\left(\sum_{q=1}^{\infty} \frac{a_q}{(n-1)^q} + \theta\right)$$

$$= a_2\left(\frac{1}{n(n-1)}\right) + a_3\left(\frac{1}{n^2} - \frac{1}{(n-1)^2}\right)$$

– **The bias of jackknife estimate $J(T)$ is in n^{-2}.**

– **Jackknife gives an estimate of the bias by:**

$$B\hat{i}as_{jack} = (n-1)(J(T) - T).$$

- **The Jackknife Variance Estimate**

$$\hat{Var}_{jack} = \frac{\sum_{j=1}^{r}(T_j^* - T)^2}{r(r-1)}.$$

  - **Monte Carlo studies that it is often conservative; that is, it often overestimates the variance (see Efron, 1982).**

- **Refer to Gentle (2002) for** *higher-order bias correction*, *the generalized jackknife*, **and** *the delete-k jackknife*.

Cross Validation and Model Selection

Cross-validation and bootstrapping are both methods for estimating generalization error based on resampling (Efron and Tibshirani, 1993).

- **Cross validation is useful in model building.**
- **In regression model building the standard problem is, given a set of potential regressors, choose a relatively small subset that provide a good fit to the data.**
  - **Standard techniques include** *stepwise* **regression and** *all best* **subsets.**
  - **If all the data are used in fitting the model, however, we have no method to validate the model.**

- A simple method to select potential regressors is to divide the sample into half, to fit the model using one half, and to check the fit using the second half.
    - The regressors to include in the model can be based on comparisons of the predictions made for the second half with the actual data.
- Instead of dividing the sample into half, we could form multiple partial data sets with overlap.
    - One way would be to leave out just one observation at a time.
    - The method of variable selection called PRESS, suggested by Allen (1971), does this. (See also Allen, 1974.)

- **Cross validation is a common method of selecting smoothing parameters.**
  - **Think of choosing window width for window estimate in regression.**
- **The resulting estimates of generalization error are often used for choosing among various models.**
- **Apparent Error and True Error:**
  Consider the problem of predicting Y using some function of X such that $E[Y-g(X)]^2$ is as well as possible.
  - **Usually, $g(X)$ is determined be a training sample (x_i, y_i)'s.**
  - **For a new point, (x_0, y_0), how well does $\hat{g}(x_0)$ match y_0?**

- Let $L(y, g)$ be a measure of the error between an observed value y and the predicted value $g(x)$. (Usually, L is the square, $[y - g(x)]^2$.)

- For prediction error,
  - Recall the residual of sum squares we learned in regression analysis.
    - Can we use RSS to do model selection?
    - RSS is typically smaller than the true error because the fit was chosen so as to minimize it.
    - RSS is the so-called apparent error.

- Define the excess error as the random variable

$$D(Y, P_{(X,Y)}) = E_{P_{(X,Y)}}[L(Y_0, \hat{g}(X_0))] - E_{\hat{P}_{(X,Y)}}[L(Y_0, \hat{g}(X_0))],$$

where $\hat{P}_{(X,Y)}$ is the estimated cumulative distribu-

tion function of (X, Y).

- **If $\hat{P}_{(X,Y)}$ is the empirical CDF the density is just $1/n$ at the sample points, so**

$$E_{\hat{P}_{(X,Y)}}[L(Y_0, \hat{g}(X_0))] = \frac{1}{n} \sum_i L(Y_i, \hat{g}(X_i)).$$

- **This quantity, which is easy to compute, is the apparent error.**

- **Cross validation methods (and other resampling methods) can be used to estimate the true error.**

Cross Validation

Consider model selection.

- In k-fold cross-validation, you divide the data into k subsets of (approximately) equal size v.
- Train the model k times, each time leaving out one of the subsets from training (estimating unknown parameters etc.), but using only the omitted subset to compute whatever chosen error criterion.
- If k equals the sample size, this is called leave-one-out.
    - Leave-one-out cross-validation is also easily confused with jackknifing since both involve omitting each training case in turn.

- Jackknifing can be used to estimate the bias of the training error and hence to estimate the generalization error, but this process is more complicated than leave-one-out cross-validation

- Leave-v-out is a more elaborate and expensive version of cross-validation that involves leaving out all possible subsets of v cases.

- For an insightful discussion of the limitations of cross-validatory choice among several learning methods, see Stone (1977).
  - Leave-one-out cross-validation often works well for estimating generalization error for continuous error functions such as the mean squared error, but it may perform poorly for discontinuous er-

ror functions such as the number of misclassified cases.

  - **In the latter case, k-fold cross-validation is preferred.**
  - **But if k gets too small, the error estimate is pessimistically biased because of the difference in training-set size between the full-sample analysis and the cross-validation analysis.**
  - **A value of 10 for k is popular for estimating generalization error.**

- **Refer to Chapter 7.11 of HTF book on bootstrap method.**

Shao (1993, $JASA$) obtained the surprising result that for selecting subsets of inputs in a linear regression, the

probability of selecting the $best$ does not converge to 1 (as the sample size n goes to infinity) for leave-v-out cross-validation unless the proportion v/n approaches 1.

- **To obtain an intuitive understanding, let's review what is generalization error.**
- **Generalization error can be broken down into three additive parts,**
  - **noise variance**
  - **estimation variance**
  - **squared estimation bias**
- **Noise variance is the same for all subsets of inputs.**
- **Bias is nonzero for subsets that are not $good$, but it's zero for all $good$ subsets, since we are assuming**

that the function to be learned is linear. Hence the generalization error of *good* **subsets will differ only in the estimation variance.**

- **The estimation variance is** $(2p/t)s^2$ **where** p **is the number of inputs in the subset,** t **is the training set size, and** s^2 **is the noise variance.**
  - **The** *best* **subset is better than other** *good* **subsets only because the** *best* **subset has (by definition) the smallest value of** p**.**
  - **But the** t **in the denominator means that differences in generalization error among the** *good* **subsets will all go to zero as** t **goes to infinity.**
  - **Therefore it is difficult to guess which subset is** *best* **based on the generalization error even when**

t **is very large.**

- **It is well known that unbiased estimates of the generalization error, such as those based on** AIC**,** FPE**, and** C_p**, do not produce consistent estimates of the** $best$ **subset (e.g., see Stone, 1979).**
- **References:**

  - **Efron, B. (1983). Estimating the error rate of a prediction rule: Improvement on cross-validation,** $JASA$ $\mathbf{78}$**, 316-331.**
  - **Efron, B. and Tibshirani, R.J. (1997). Improvements on cross-validation: The** $.632+$ **bootstrap method.** $JASA$ $\mathbf{92}$**, 548-560.**
  - **Stone, M. (1977). Asymptotics for and against cross-validation.** $Biometrika$ $\mathbf{64}$**, 29-35.**