The upstrap paper vignette

Ciprian M. Crainiceanu and Adina Crainiceanu

This is a vignette associated with the paper "The upstrap". Just like the bootstrap, the upstrap uses sampling with replacement of the original data to estimate the variability of various estimators (a.k.a. functions of the data). In contrast to the bootstrap, the upstrap uses a number of resamples that can be smaller than, equal to, or larger than the original sample size of the data; the bootstrap uses only a number of samples equal to the sample size of the original data. This substantially extends the number and variety of scenarios where the variability of estimators can be evaluated using computational tools. For example, the upstrap could be used to assess the low and moderate sample size behavior of estimators via subsampling with replacement. Also, it can be used for sample size calculations that take into account the variability in effect estimation. To illustrate this latter point, this vignette provides an example on how to conduct the upstrap when we are interested in sample size calculations in a regression context. This document is designed to make the paper fully reproducible, while an `R` (R Development Core Team, 2008) package called `upstrap` will be deployed at the same GitHub address. The vignette will be improved, as software is updated and examples are added. The code is presented in extended format for pedagogical purposes, but the code can be substantially reduced using the (still in development) `upstrap` package in `R`.

To do this, we will use a subset of the Sleep Heart Health Study (SHHS), a multicenter study on sleep-disordered breathing, hypertension, and cardiovascular disease (Dean et al. 2016, Quan et al. 1997, Redline et al. 1998). The SHHS drew on the resources of existing, well-characterized, epidemiologic cohorts, and conducted further data collection, including measurements of sleep and breathing. The complete SHHS dataset is available online as part of the National Sleep Research Resource https://sleepdata.org/datasets. For the purpose of this vignette we consider a small subset of the variables from the visit 1 of the SHHS. We use `R` for computation and visualization.

Read in the SHHS data

For ilustration purposes we analyze a subset of the SHHS visit 1 data, which contain 5804 participants and 29 variables plus the patient id column. This subset was created for the book Methods in Biostatistics with R (Crainiceanu, Caffo, Muschelli, 2018). Data are provided with this vignette, making the document and the associated paper reproducible. We start by setting the path for where the data are stored on the computer.

```
# set the working directory
# this will need to be changed on a different computer
f.path<-"path_to_upstrap_code_folder"
setwd(f.path)
```

Data are read into the list `dat`

```
# read in the data
dat = read.table(file = "data/shhs1.txt",
                 header = TRUE, na.strings="NA")
```

We show the head of the list `dat`

```
head(dat)
```

```
##   pptid waist COPD15 ASTHMA15 slp_lat time_bed timest1p timest2p times34p
## 1     1    86      0        0      NA    440.5 6.258322 60.85220 19.30759
## 2     2   107      0        0      NA    225.0 0.824176 65.65934 16.75824
## 3     3    82      0        0      NA    431.5 4.881451 40.30683 42.81730
## 4     4    85      0        0    14.0    358.5 2.990033 29.40199 52.32558
## 5     5    76      0        0     6.5    477.0 5.675676 68.64865 13.37838
```

```
## 6     6    95      0        0      NA    469.5 6.201550 68.34625  5.03876
##   timeremp      rdi4p StLOutP StOnsetP SlpPrdP Staging1 Staging2 Staging3
## 1 13.58189  1.4380826      28       28   375.5        0        0        0
## 2 16.75824 17.8021978       0        0   182.0       NA       NA       NA
## 3 11.99442  4.8535565     167      167   358.5        1        0        0
## 4 15.28239  0.7973422      54       82   301.0        1        0        0
## 5 12.29730  2.7567568       7       20   370.0        1        0        0
## 6 20.41344  3.7209302     123      123   387.0        1        0        0
##   Staging4 Staging5 RestAn1 RestAn2 RestAn3 RestAn4 HTNDerv_s1 shhs1_tcvd
## 1        1        0       1       0       0       0          1          0
## 2       NA       NA      NA      NA      NA      NA          1          0
## 3        0        0       1       0       0       0          0          0
## 4        0        0       1       0       0       0          1          0
## 5        0        0       1       0       0       0          1          0
## 6        0        0       1       0       0       0          1          0
##   gender age_s1 smokstat_s1  WASO   bmi_s1
## 1      1     55           2  65.0 21.77755
## 2      1     78           0  43.0 32.95068
## 3      0     77           0  73.0 24.11415
## 4      1     48           0  43.5 20.18519
## 5      0     66           2 100.5 23.30905
## 6      1     63           0  82.5 27.15271
```

and the dimension of the list

```
dim(dat)
```

```
## [1] 5804   30
```

The dataset contains 5804 rows and 30 columns. Each row corresponds to a person enrolled in the SHHS and each column corresponds to a person-specific variable. The only exception is the first column labeled `pptid`, which is the subject identifier. For the purpose of this document we will only use the variables `gender`, `age_s1`, `bmi_s1`, `HTNDerv_s1`, and `rdi4p`. The variable `gender` is self explanatory and the sample included 52.4% women and 47.6% men. The variables `age_s1` and `bmi_s1` are the age and Body Mass Index (BMI) at visit 1 of the SHHS. The visit number is the reason for the underscore `_s1` on these variables. The variable `HTNDerv_s1` is whether (coded as 1) or not (coded as a 0) the person has a blood pressure (BP) greater than or equal to 140/90 mmHg or they are under current treatment with anti-hypertensive medication. The variable `rdi4p` is the the overall respiratory disturbance index (RDI) at 4% oxygen desaturation. This is the ratio of the count of all apneas and hypopneas associated with at least a 4% oxygen desaturation to the total sleep time expressed in hours (https://sleepdata.org/). The 4% oxygen desaturation refers to blood's oxygen level drop by 4% from baseline. This variable is often used to characterize the severity of sleep apnea, with larger values corresponding to worse outcomes. Some of the other covariates that will not be used in our vignette are self explanatory, though for more in-depth description of the covariates a simple google search on the name of the variable and SHHS should provide the necessary information. Having a complete dictionary of variables is not necessary for this document.

To explore the data, we first provide a scatterplot of the Body Mass Index (BMI) versus Respiratory Disturbance Index (RDI), variables labeled `bmi_s1` and `rdi4p`, respectively. The figure below provides the scatter plot of BMI versus RDI, though, for presentation purposes, we only show RDI less than 50. The plot indicates that RDI$\geq$ 0 (by definition), that there is a higher density of observations for small RDI (note the deeper shades of blue close to the x-axis) and that most people in the SHHS sample have a BMI$\geq$ 20. The figure indicates that there is a large amount of variability, as for any BMI level there is a large range of observed RDI values.

We also calculate the mean BMI and RDI

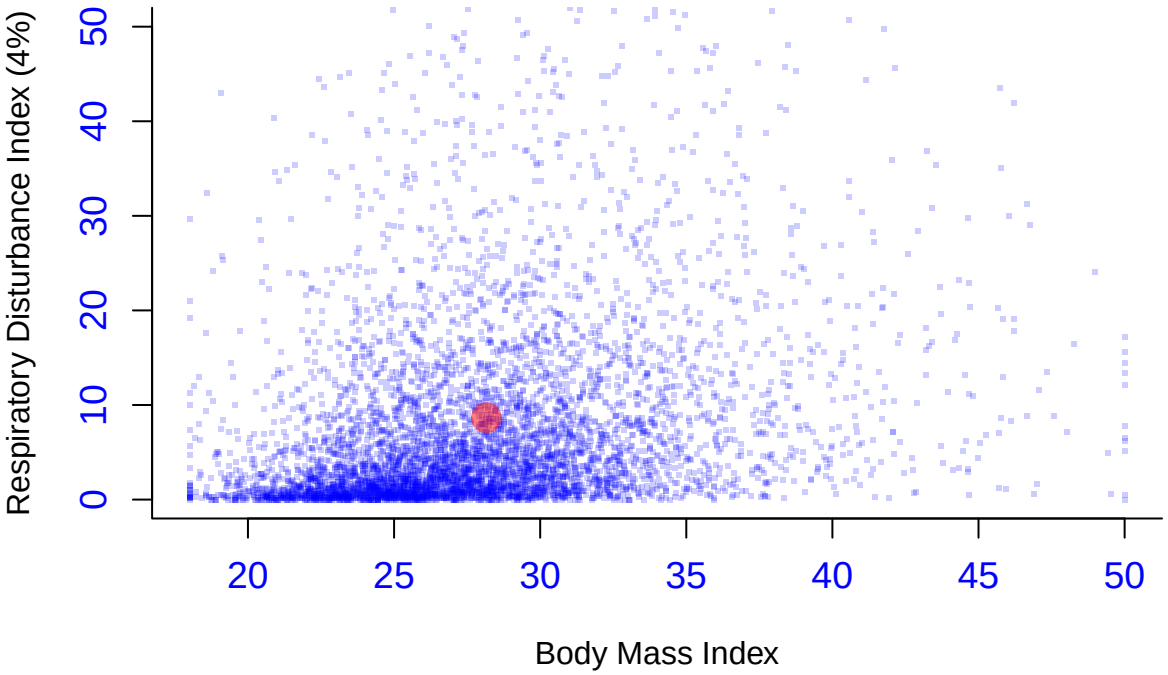


Figure 1: Scatter plot of BMI versus RDI in the SHHS.

```
round(mean(dat$bmi_s1,na.rm=TRUE),digits=2)
```

```
## [1] 28.16
```

```
round(mean(dat$rdi4p,na.rm=TRUE),digits=2)
```

```
## [1] 8.66
```

The center of the red circle has the x-axis coordinate equal to the empirical mean BMI (28.16) and y-axis coordinate equal to the empirical mean RDI (8.66). This is the mean of the two dimensional vector of observations $(\text{BMI}_i, \text{RDI}_i)$ for $i = 1, \ldots, 5804$.

Fit a basic regression model

Consider the case of a regression problem where the outcome is binary and we have some main effects and some interactions. The outcome in our case is moderate to severe sleep apnea, which is defined as an `rdi4p` at or above 15 events per hour. Using the code below

```
# define the moderate to severe sleep apnea variable from rdi4p
MtS_SA=dat$rdi4p>=15

# identify how many individuals in SHHS (visit 1) have moderate
# to severe sleep apnea
n_positive =sum(MtS_SA)
n_positive
```

```
## [1] 1004
```

we obtain that there are 1004 individuals in the SHHS who have moderate to severe sleep apnea among the individuals in SHHS for a prevalence of 17.3% of the SHHS population. Individuals in SHHS were oversampled for increased likelihood of having sleep apnea and this prevalence is not representative for the

overall US population. Indeed, the prevalence of obstructive sleep apnea associated with accompanying daytime sleepiness is approximately 3 to 7% for adult men and 2 to 5% for adult women in the general population (Punjabi 2008). If Y_i denotes the moderate to severe sleep apnea variable for subject i then we model $Y_i \sim \text{Bernoulli}(p_i)$, where

$$\text{logit}(p_i) = \beta_0 + \beta_1 \text{sex}_i + \beta_2 \text{age}_i + \beta_3 \text{BMI}_i + \beta_4 \text{HTN}_i + \beta_5 \text{age}_i \times \text{HTN}_i \, .$$

Here for subject i, we used the labels sex_i for `gender`, age_i for `age_s1`, BMI_i for `bmi_s1`, and HTN_i for `HTNDerv_s1`. The model is simple and contains main effects of sex_i, age_i, BMI_i, and HTN_i and an interaction between age_i and HTN_i. To do this, we fit a generalized linear model (GLM) regression, which is done in `R` as

```
# conduct a binary GLM regression with outcome moderate to severe
# sleep apnea (Yes=1, No=0) on a few covariates (main effects)
# and an interaction (age by HTN)
fit<-glm(MtS_SA~gender+age_s1+bmi_s1+HTNDerv_s1+age_s1*HTNDerv_s1,
           family="binomial",data=dat)
summary(fit)
```

```
##
## Call:
## glm(formula = MtS_SA ~ gender + age_s1 + bmi_s1 + HTNDerv_s1 +
##     age_s1 * HTNDerv_s1, family = "binomial", data = dat)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0443  -0.6559  -0.4441  -0.2671   2.8556
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -8.710927   0.421770 -20.653  < 2e-16 ***
## gender             1.156897   0.079181  14.611  < 2e-16 ***
## age_s1             0.036892   0.005000   7.378 1.61e-13 ***
## bmi_s1             0.138369   0.007412  18.669  < 2e-16 ***
## HTNDerv_s1         0.758377   0.473129   1.603    0.109
## age_s1:HTNDerv_s1 -0.008748   0.007146  -1.224    0.221
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5314.5  on 5760  degrees of freedom
## Residual deviance: 4659.1  on 5755  degrees of freedom
##   (43 observations deleted due to missingness)
## AIC: 4671.1
##
## Number of Fisher Scoring iterations: 5
```

There are several important findings in the output, but there is nothing non-basic about it. Therefore, we will not discuss the output in detail, as this document is not designed as a complete case study of SHHS. However, we will focus on the point estimator of the main effect of HTN (`HTNDerv_s1`) based on the entire data set. The realization of this estimator based on the SHHS is 0.758 with a p-value of 0.109. We would like to answer the following question: "For a given value of β (say, 0.1 or 0.2), what is the sample size that will ensure that the null hypothesis of zero main effect of HTN will be rejected with a frequency at least equal to $100(1-\beta)\%$ (power)?" Here we do not discuss whether this is a relevant scientific question, just providing an example of a question that would be very hard to answer using other methods. The techniques described below apply exactly the same for any model parameter.

The upstrap for sample size calculation

To answer the question we set a grid of sample sizes and for each sample size we upstrap the data (sample with replacement) with that particular sample size. Because the p-value is larger than 0.05 in the original SHHS dataset we expect that the sample size necessary to achieve $> 80\%$ power is larger than the original sample size. Therefore, we consider sample sizes that multiply the original sample size by $1, 1.2, \ldots, 5$. For example, for the grid point 1.2 it means that we sample with replacement $1.2 \times 5804 = 6964.8$ subjects. Because this is not an integer the implementation of the `sample` function in `R` rounds it down to 6964 subjects. For grid point 1.2 we upstrap the data with a sample size 6964 for a given number of times, `n.upstrap`; here we chose this number to be equal to 10000 to eliminate some of the sampling variability. For somebody who is familiar with the bootstrap the upstrap should be straightforward to understand. It has become accepted that the bootstrap should sample the same number of subjects with the number of subjects in the original dataset. Here we argue that sometimes one may be interested in sampling more or fewer subjects than in the original dataset.

For every grid point we calculate the percent of times (out of the total number of upstraps for that grid point) we reject the null that the HTN effect is equal to zero. Below we provide the commented code to do that. The code was run separately and we use (`eval=FALSE`) to avoid running 420000 Bernoulli regressions every time the Rmarkdown document is compiled.

We start by setting up some of the basic ingredients for the upstrap including the original sample size, labeled `n.oss`, the grid of multipliers for the original sample size, labeled `multiplier.grid`, and the number of upstraps per point in the multiplier grid, `n.upstrap`.

```
# set the seed for reproducibility
set.seed(08132018)

# sample size of the original data
n.oss=dim(dat)[1]

# number of grid points for the multiplier of the sample size
n.grid.multiplier=21

# minimum and maximum multiplier for the sample size
# here they are set to 1 (same sample size) and 5 (5 times the original sample size)
min.multiplier=1
max.multiplier=5

# set the grid of multipliers for the original sample size
# here 1.2 stands for a sampel size that is 20% larger than the original sample size
multiplier.grid=seq(from=min.multiplier, to=max.multiplier,length=n.grid.multiplier)

# set the number of upstraps for each grid point
n.upstrap=10000
```

Once these variables are set up, we set up the matrix `check` of dimensions `n.upstrap` by `n.grid.multiplier`. The (u, r) entry of this matrix records whether the p-value of the HTN variable for the rth multiplier of the sample size on the uth upstrap sample is less than 0.05. One could record the p-values instead and compute the `check` matrix at the end, but this does not change the general idea.

```
# build the matrix that will contain whether or not the null hypothesis that the
# HTN is zero for a given multiplier (r) and upstrap sample (u)
check<-matrix(rep(NA,n.upstrap*n.grid.multiplier),ncol=n.grid.multiplier)

# here j is the index of the multiplier of the original sample size, n.oss
for (j in 1:n.grid.multiplier)
```

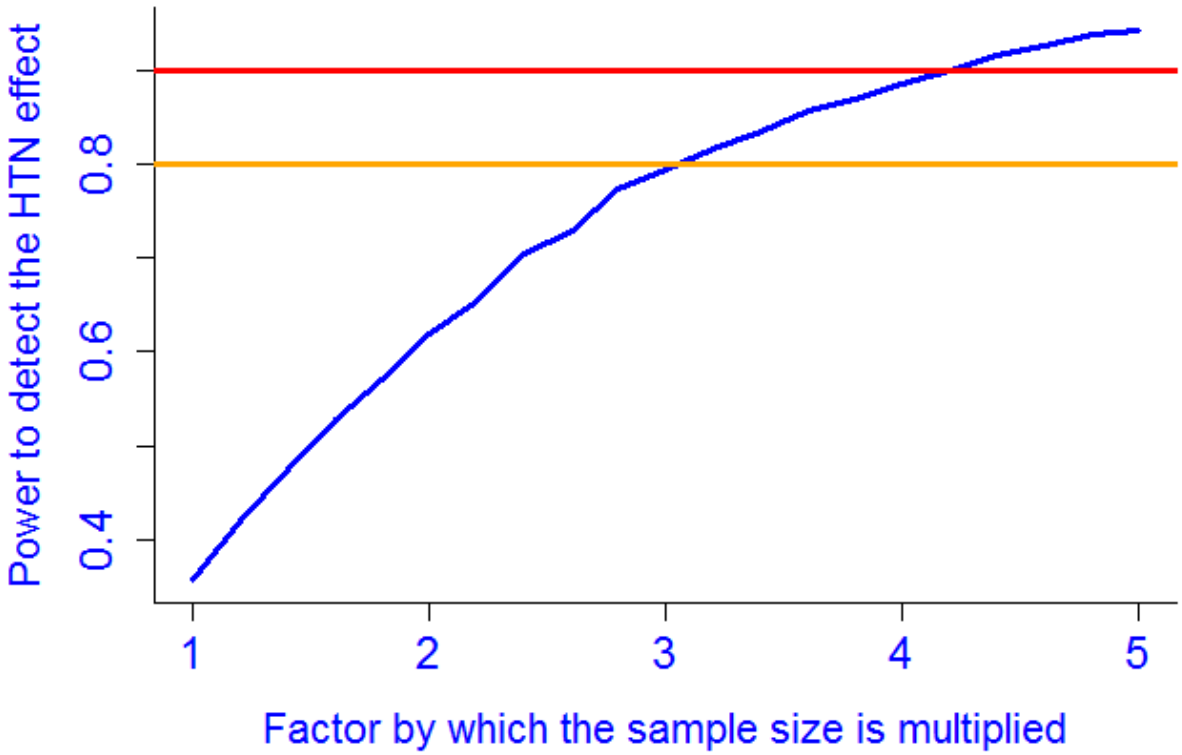
```
  {#Each loop corresponds to an increase in the sample size

  # here u is the u-th upstrap for the r-th sample size multiplier
  # this simulation can/will be done more efficiently using matrices
  for (u in 1:n.upstrap)
  {# each loop corresponds to an upstrap with a given sample size

  # construct the upstrap sample index
      # from 1, ..., n.oss (original sample size)
      # size equal to original sample size times the multiplier
      # sampling is with replacement
  temp_index<-sample(1:n.oss,n.oss*multiplier.grid[j],replace=TRUE)

  # extract the data (covariates and outcome) using the upstrap sample index
  temp_data<-dat[temp_index,]
  MtS_SA=temp_data$rdi4p>=15

  # fit the same model on the upstrapped data
  fit<-glm(MtS_SA~gender+age_s1+bmi_s1+HTNDerv_s1+age_s1*HTNDerv_s1,
           family="binomial",data=temp_data)

  # obtain the p-value for HTN in the upstrapped data
  check[u,j]<-coef(summary(fit))[,4][5]<0.05
  }
}
```

Once the matrix `check` of dimensions `n.upstrap` by `n.grid.multiplier` is obtained we calculate its column means. This results in a vector of length `n.grid.multiplier`, which contains the proportion of times out of `n.upstrap` samples when the null hypothesis that the parameter of HTN is equal to zero is rejected. This is the power curve as a function of the multiplier of the original sample size.

```
# power check calculates the proportion of times the null hypothesis is rejected
power_check<-colMeans(check)
```

Below we plot the vector of multipliers of the original sample size, `n.grid.multiplier`, versus the power curve for rejecting the null hypothesis that the parameter of HTN is equal to zero, `power_check`. Horizontal lines are shown to indicate power equal to 0.8 (orange) and 0.9 (red). The value 1 on the x-axis corresponds to a sample size equal to the original sample size. The corresponding value on the y-axis on the curve is the boostrap p-value, or the frequency with which the null hypothesis that the HTN effect is zero is rejected using the bootstrap. The value 1.2 on the x-axis corresponds to a sample size 20% larger than the original sample size. The corresponding value on the y-axis on the curve is the upstrap p-value, or the frequency with which the null hypothesis that the HTN effect is zero is rejected using the upstrap with a sample size multiplier equal to 1.2.

```
# plot the vector of multipliers of the original sample size, versus the power curve for rejecting the
plot(multiplier.grid,power_check,type="l",col="blue",lwd=3,
     xlab="Factor by which the sample size is multiplied",
     ylab="Power to detect the HTN effect",
     bty="l",cex.axis=1.5,cex.lab=1.4,col.lab="blue",
     col.axis="blue",main=NULL)
# add horizontal lines to indicate power equal to 0.8 (orange) and 0.9 (red).
abline(h=0.8,lwd=3,col="orange")
abline(h=0.9,lwd=3,col="red")
```



The figure provides the frequency with which the test for no HTN effect is rejected in the model as a function of the multiplier of the sample size. For example, for the multiplier 2 we produced 10000 (upstrap) samples with replacement from the SHHS with twice the number of subjects $2n = 11608$. For each upstrapped data we ran the model and recorded whether the p-value for HTN was smaller than 0.05. At this value of the sample size we obtained that the null hypothesis of no HTN effect was rejected in 62% of the upstrap samples. We also obtained that the power was equal to 0.794 at the sample size mutiplier 3.0 and 0.817 at multiplier 3.2, indicating that the power 0.8 would be attained at $3.1 * n \approx 17,500$.

There are very few methods to estimate the sample size in such examples and we contend that the upstrap is a powerful and general method to conduct such calculations. Similar approaches could be used in many other situations, including estimating a fixed effect (e.g. treatment) using longitudinal data in the context of a clinical trial or the sample size necessary to detect gene by gene and gene by environment interactions in genomics studies.

References

1. Crainiceanu CM, Caffo B, Muschelli J. Methods in Biostatistics with R. Leanpub 2018, Leanpub link
2. Dean DA, Goldberger AL, Mueller R, et al. Scaling Up Scientific Discovery in Sleep Medicine: The National Sleep Research Resource. Sleep. 2016;39(5):1151-1164.
3. Quan SF, Howard BV, Iber C, et al. The Sleep Heart Health Study: design, rationale, and methods. Sleep. 1997;20(12):1077-1085.

4. R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

5. Punjabi NM. The Epidemiology of Adult Obstructive Sleep Apnea. Proceedings of the American Thoracic Society. 2008;5(2):136-143.

6. Redline S, Sanders MH, Lind BK, et al. Methods for obtaining and analyzing unattended polysomnography data for a multicenter study. Sleep Heart Health Research Group. Sleep. 1998;21(7):759-767.