STAT 641: BOOTSTRAPPING METHODS

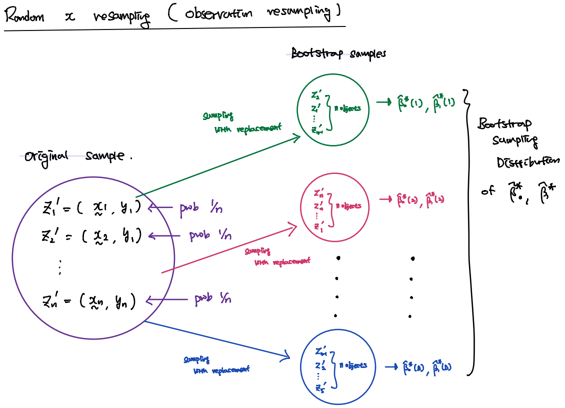
Jiyoun Myung

Department of Statistics and Biostatistics
California State University, East Bay

Spring 2021, Day 10

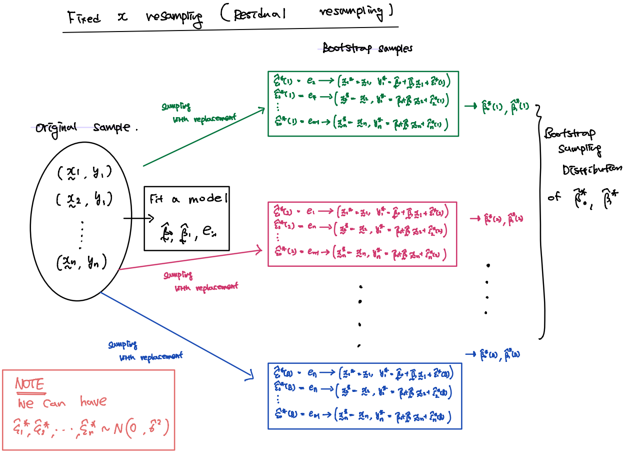
Review

Random x resampling (Obervation resampling)



Review

Fixed x resampling (Residual resampling)



Comparison

Resampling	Obervations	Residuals
Model-dependent		
Fixed design (X)		
Maintains (X, Y) association		

Differences are obvious when the regression model or data is peculiar or if there is a severe outlier.

Example

Florida 2000 US Presidential election results

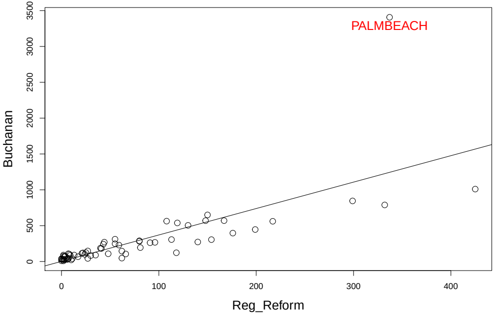
```r
# "County by county returns for the 2000 US Presidential election."
fl <- read.table("florida2000.txt", header = TRUE)
names(fl)
```

```
##  [1] "County"      "Gore"        "Bush"        "Buchanan"    "Nader"
##  [6] "Total_Votes" "Reg_Reform"  "Reg_Rep"     "Reg_Ind"     "Reg_Grn"
## [11] "Reg_Dem"     "Total_Reg"
```

Data show by county:

- predictor: number of people registered to Reform Party.
- response: number of votes received by Buchanan.

Florida 2000 US Presidential election results



Florida 2000 US Presidential election results

Slope Estimate and estimated standard error

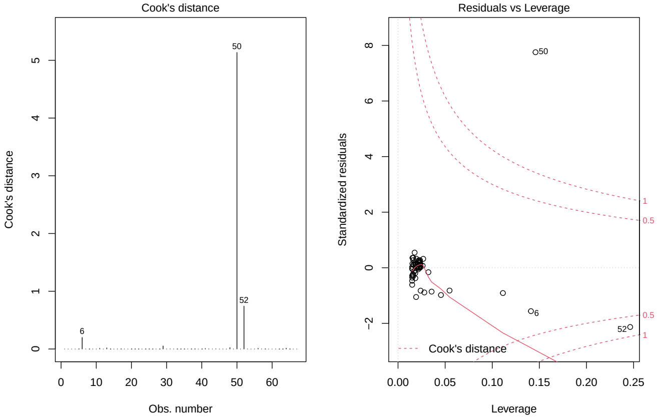
```
x <- fl$Reg_Reform
y <- fl$Buchanan
fit <- lm(y ~ x)
round(coef(summary(fit))[2,], 2)
```

```
##   Estimate Std. Error    t value   Pr(>|t|)
##       3.69       0.41       9.02       0.00
```

Florida 2000 US Presidential election results

Leverage and influential points

```
par(mfrow= c(1, 2))
plot(fit, 5:4)
```



Palm Beach is not so leveraged, but is "influential."

Florida 2000 US Presidential election results

Observation Resampling vs Residual Resampling

Observation Resampling

```
##
## ORDINARY NONPARAMETRIC BOOTSTRAP
##
##
## Call:
## boot(data = fl, statistic = boot.fl, R = 5000)
##
##
## Bootstrap Statistics :
##     original       bias    std. error
## t1* 1.532519  0.56700841   47.652449
## t2* 3.686713 -0.02172581    1.158437
```

Residual Resampling

```
##
## ORDINARY NONPARAMETRIC BOOTSTRAP
##
##
## Call:
## boot(data = fl, statistic = boot.fl.fixed, R = 500
##
##
## Bootstrap Statistics :
##     original        bias    std. error
## t1* 1.532519 -0.049059902  45.4159421
## t2* 3.686713  0.000824195   0.3903628
```

Florida 2000 US Presidential election results

R code for previous page

Observation Resampling

```
library(boot)

boot.fl <- function(data, indices){
   # select obs. in bootstrap sample
  data <- data[indices,]
  mod <- lm(Buchanan ~ Reg_Reform, data = data)
  # return coefficient vector
  coefficients(mod)
}

fl.boot <- boot(fl, boot.fl, 5000)
fl.boot
```

Residual Resampling

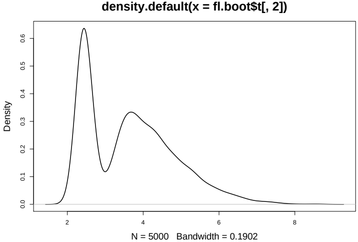
```
fits <- fitted(fit)
e <- residuals(fit)
X <- model.matrix(fit)

boot.fl.fixed = function(data, indices) {
  y_b <-  fits + e[indices]
  mod <- lm(y_b ~ X - 1)
  coefficients(mod)
}

fl.fixed.boot <- boot(fl, boot.fl.fixed, 5000)
fl.fixed.boot
```

Observation Resampling vs Residual Resampling

Observation Resampling vs Residual Resampling

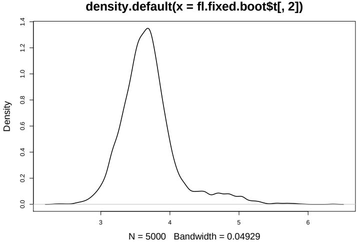
Observation Resampling

```
plot(density(fl.boot$t[,2]))
```



Residual Resampling

```
plot(density(fl.fixed.boot$t[,2]))
```



Which bootstrap method is better? The answer depends on how far we trust the linear regression model.

Observation Resampling vs Residual Resampling

- **Observation resampling** is a good choice when we are modeling observational data in which the explanatory variables are observed randomly from a population.
- **Residual resampling** is a good choice if we are analyzing data from a designed experiment in which the explanatory variables have a small number of specified values.
- Residual resampling requires a "true" model in order to obtain the residuals which are resampled. Observation (or random) resampling does not. Residual resampling keeps the same X's in every bootstrap sample.
- As the sample size grows (with other conditions), two methods become similar, assuming the model is correctly identified.
- Random resampling usually leads to a larger estimate of standard error (with enough bootstrap replications) since it allows for more sources of variation (from randomness in X's).
- Bootstrap SE of residual resampling will be close to classical SE (OLS formula) as $B \to \infty$. But, Bootstrap SE of observation resampling does not always agree with classical SE.

Your Turn

Dataset **catsM** contains a set of data on the heart weights and body weights of 97 male cats. We investigate the dependence of **heart weight** (g) on **body weight** (kg). The data set is available in the boot package.

(a) Investigate the data set by first fitting a straight line regression and creating diagnostic plots.

(b) Next, perform model-based bootstrap regression (residual resampling). Are the bootstrap estimates for intercept and slopes appear normal? Is the model-based standard error for the original fit accurate?

(c) Do you think the results are effected by any single observation?

(d) Perform the observation resampling method. And compare the results with (b) and (c).