Bootstrap in Linear Models:
a comprehensive R package

Megan Heyman, PhD
heyman@rose-hulman.edu
May 16, 2019

ROSE-HULMAN
INSTITUTE OF TECHNOLOGY

Traditional Linear Model

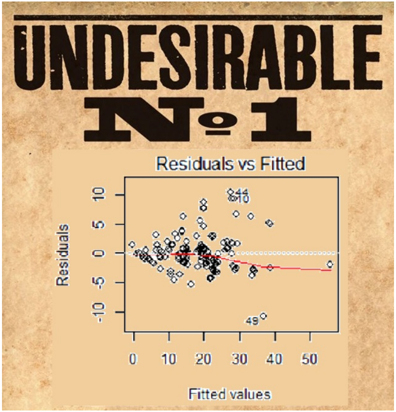
Suppose you want to model a response, $\boldsymbol{Y}$, with a linear model in a set of predictors, $\mathcal{X}$:

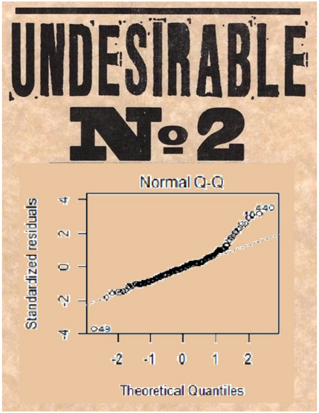
$$\boldsymbol{Y} = \mathcal{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

(Very) Traditional Assumptions:

- $\mathcal{X}$ is of full column rank
- The errors ($\boldsymbol{\epsilon}$) are independent.
- $\mathrm{E}(\boldsymbol{\epsilon}) = \mathbf{0}$.
- ...

Residual Plots

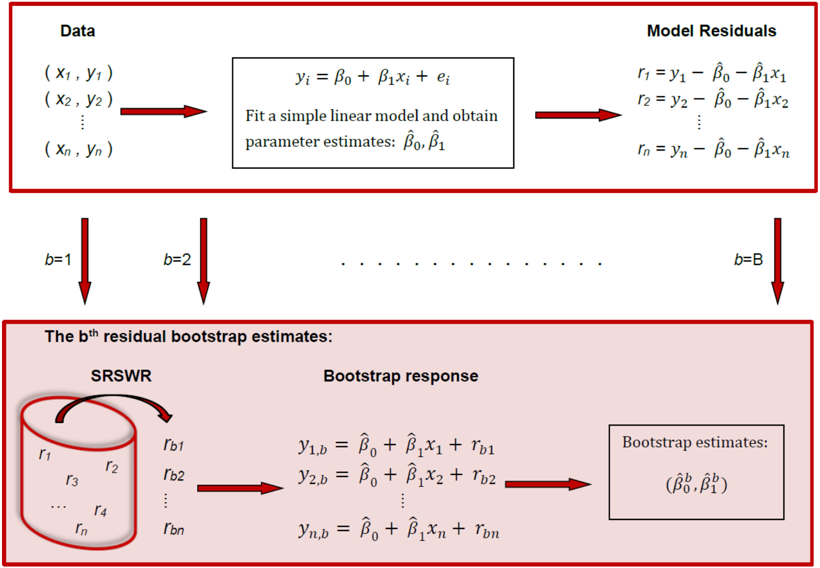




Outline

1. Review Bootstrap in Multiple Linear Regression
    a. Residual
    b. Paired
    c. Wild
2. Classification of Bootstrap Methods
3. Bootstrap in **R** for Linear Models
4. Example Data Analysis

Residual Bootstrap in Simple Linear Regression



ROSE-HULMAN INSTITUTE OF TECHNOLOGY

Residual Bootstrap in Multiple Linear Regression

$$W_i = (w_{i,1}, w_{i,2}, ..., w_{i,n}) \sim \textsf{Multinomial}(1, 1/n, 1/n, ..., 1/n)$$

For the b^{th} bootstrap sample, generate $i = 1, 2, ..., n$ independent W_i to compose the matrix $\mathcal{W}_b$.

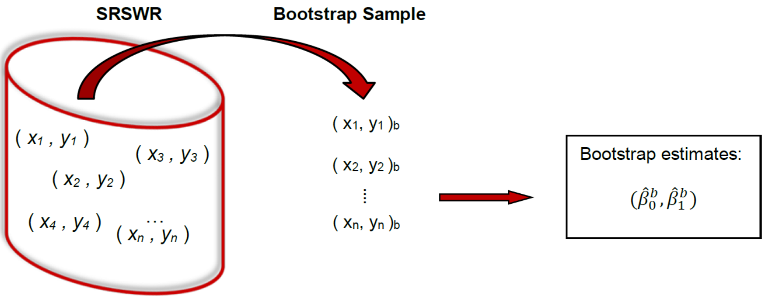
$$\mathcal{W}_b = \begin{bmatrix} w_{1,1} & w_{1,2} & \dots & w_{1,n} \\ w_{2,1} & w_{2,2} & \dots & w_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{n,1} & w_{n,2} & \dots & w_{n,n} \end{bmatrix}$$

The residual bootstrap estimate for $\hat{\boldsymbol{\beta}}$ is

$$\hat{\boldsymbol{\beta}}_b = (\mathcal{X}^T\mathcal{X})^{-1}\mathcal{X}^T\mathbf{Y}_b = \hat{\boldsymbol{\beta}} + (\mathcal{X}^T\mathcal{X})^{-1}\mathcal{X}^T\mathcal{W}_b\boldsymbol{r}$$

Paired Bootstrap in Simple Linear Regression

What about just resampling from the original observations?



This resampling scheme is called the paired bootstrap.

Paired Bootstrap in Multiple Linear Regression

- Similarly to residual bootstrap, multinomial weights on the rows of $\mathcal{X}$ and $\boldsymbol{Y}$.
- The bootstrap estimator is similar to the weighted least squares estimator (Chatterjee, 2000)

The b^{th} paired bootstrap estimator for $\hat{\boldsymbol{\beta}}$ is

$$\hat{\boldsymbol{\beta}}_b = (\mathcal{X}^T \mathcal{W}_b \mathcal{X})^{-1} \mathcal{X}^T \mathcal{W}_b \boldsymbol{Y}$$

Unfortunately, this straight-forward estimation technique is computationally expensive.

Motivation for the Wild Bootstrap

The $w_{i,k}$ for $k = 1, 2, ..., n$ are not independent, but form a simplex. The dependence makes higher order moment calculations difficult for residual and paired bootstrap.

What about weighting the residuals by independent random variables?

- When the intercept is in a linear model, $\sum_{i=1}^{n} r_i = 0$.
- If $E(\epsilon) = 0$ then the residuals vs. fits should not have a systematic pattern around 0.

$\Rightarrow$ Perturbing residuals by a random variable with 0 mean essentially emulates the variability associated with sampling.

Wild Bootstrap in Simple Linear Regression

Data

(x_1, y_1)
(x_2, y_2)
$\vdots$
(x_n, y_n)

$$y_i = \beta_0 + \beta_1 x_i + e_i$$

Fit a simple linear model and obtain parameter estimates: $\hat{\beta}_0, \hat{\beta}_1$

Model Residuals

$r_1 = y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1$
$r_2 = y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2$
$\vdots$
$r_n = y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n$

b=1 b=2 . . . . b=B

The b^{th} wild bootstrap estimates:

Indep. R.V.s:

$U_{1,b}$
$U_{2,b}$
$\vdots$
$U_{n,b}$

Bootstrap Response

$y_{1,b} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + U_{1,b}\, r_1$
$y_{2,b} = \hat{\beta}_0 + \hat{\beta}_1 x_2 + U_{2,b}\, r_2$
$\vdots$
$y_{n,b} = \hat{\beta}_0 + \hat{\beta}_1 x_n + U_{n,b}\, r_n$

Bootstrap estimates:

$(\hat{\beta}_0^b, \hat{\beta}_1^b)$

Wild Bootstrap in Multiple Linear Regression

Let U_i for $i = 1, 2, ..., n$ be independent random variables with $E(U_i) = 0$ and $Var(U_i) = 1$.

Then, the **wild bootstrap** weights the estimated model residuals with $\mathcal{U}_b$.

$$\mathcal{U}_b = \text{diag}(U_1, U_2, ..., U_n)$$

The b^{th} wild bootstrap estimator for $\hat{\boldsymbol{\beta}}$ is

$$\hat{\boldsymbol{\beta}}_b = (\mathcal{X}^T\mathcal{X})^{-1}\mathcal{X}^T\boldsymbol{Y}_b = \hat{\boldsymbol{\beta}} + (\mathcal{X}^T\mathcal{X})^{-1}\mathcal{X}^T\mathcal{U}_b\boldsymbol{r}$$

ROSE-HULMAN INSTITUTE OF TECHNOLOGY

Which Bootstrap Method Should You Use?

For the linear model, (Liu and Singh, 1992) showed that all bootstrap methods may be classified as either **Efficient** or **Robust**. These terms are related to the variance of the estimators.

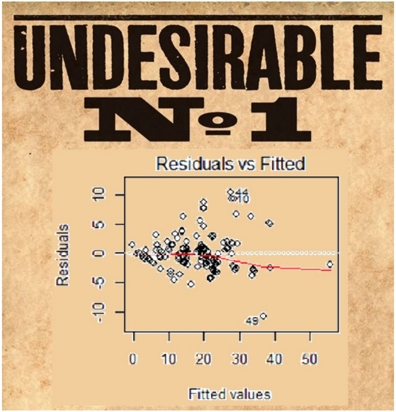
Efficient: Estimator has additional efficiency

- Must assume errors are independent, have mean 0, and have constant variance.

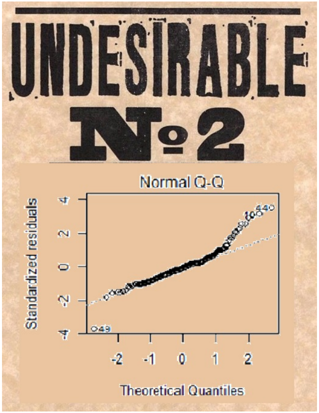
Robust: Estimator is $\sqrt{n}$-consistent with heteroscedastic errors

- Must assume errors have mean 0 and are independent.

RE: Residual Plots

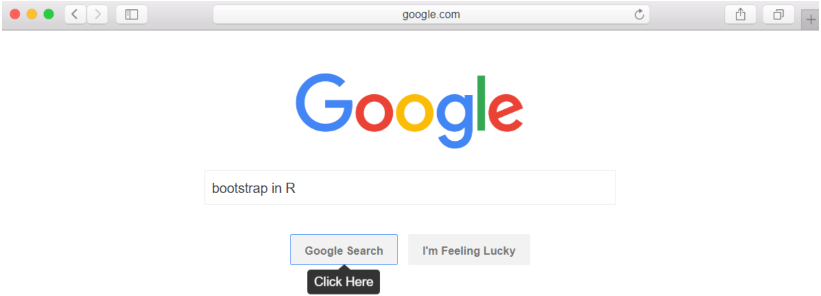


Robust bootstrap: wild or paired



Efficient bootstrap: residual

Ready to use bootstrap in R...



Well-Known Functions (Top Google Results)

`boot(data= , statistic= , R=, ...)`

- Function inside the `boot` package
- Users must specify the function/statistic to bootstrap
- Allows for bootstrap within strata

`bootstrap(x, nboot, theta, ..., func=NULL)`

- Function inside the `bootstrap` package
- Seems to have same or less flexibility than `boot()`

```
Boot(object, f=coef, labels=names(f(object)), R=999,
    method=c("case", "residual"), ncores=1, ...)
```

- Function inside the `car` package
- Easier for novice **R** users to implement
- Only implements paired or residual bootstrap types

Searching R

```
library(packagefinder)
findPackage("bootstrap")
Results:  286 out of 14208 CRAN packages found in 4
seconds...
```

SCORE	NAME	DESC_SHORT	GO
100.0	hcci	Interval estimation for the parameters of linear models with heteroskedasticity (Wild Bootstrap)	5130
100.0	shinybootstrap2	Bootstrap 2 Web Components for Use with Shiny	11738
87.5	Omisc	Univariate Bootstrapping and Other Things	8308
87.5	WiSEBoot	Wild Scale-Enhanced Bootstrap	13975
75.0	bsplus	Adds Functionality to the R Markdown + Shiny Bootstrap Framework	1281
62.5	bbw	Blocked Weighted Bootstrap	774
62.5	bootstrap	Functions for the Book "An Introduction to the Bootstrap"	1186
62.5	bootstrapFP	Bootstrap Algorithms for Finite Population Inference	1187
62.5	bootSVD	Fast, Exact Bootstrap Principal Component Analysis for High Dimensional Data	1189
62.5	geotoolsR	Tools to Improve the Use of Geostatistic	4523
62.5	knitrBootstrap	'knitr' Bootstrap Framework	6110
50.0	BaBooN	Bayesian Bootstrap Predictive Mean Matching - Multiple and Single Imputation for Discrete Data	590
50.0	boot	Bootstrap Functions (Originally by Angelo Canty for S)	1174
50.0	boot	Bootstrap Functions (Originally by Angelo Canty for S)	14194
50.0	bootsPLS	Bootstrap Subsamplings of Sparse Partial Least Squares - Discriminant Analysis for Classification and Signature Identification	1184

Putting it all together: `lmboot`

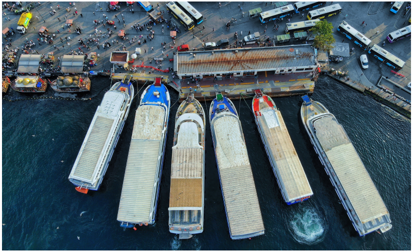
Objective:
A comprehensive R package which implements various bootstrap techniques in linear models. Straight-forward implementation is highly preferred for novice **R** users.

```
library(lmboot)
```

Current functions for bootstrapping in `lm()` models:
`residual.boot()`: Residual bootstrap
`paired.boot()`: Paired bootstrap
`wild.boot()`: Wild bootstrap
`jackknife()`: Delete-1 Jackknife
`bayesian.boot()`: Bayesian bootstrap
`ANOVA.boot()`: Bootstrap in 2-way ANOVA (wild and residual)

ROSE-HULMAN INSTITUTE OF TECHNOLOGY

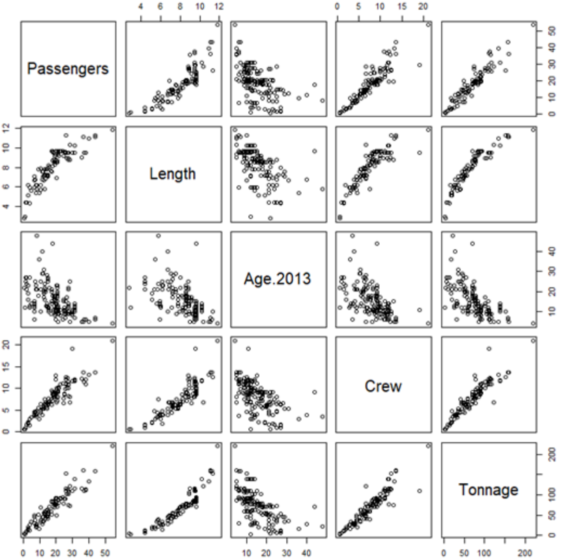
Example: Cruise Ship Properties



http://www.stat.ufl.edu/~winner/data/cruise_ship.txt
contains information about 158 cruise ships as of 2013.
Let's consider modeling the number of passengers based on the ship age, tonnage, length, and crew size.

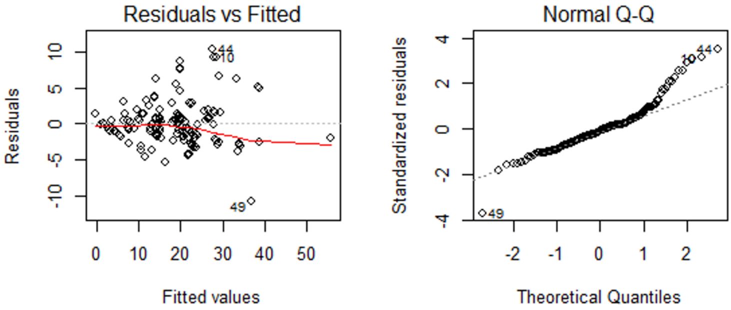
Example: Cruise Ship Properties

```
pairs(~Passengers+Length+Age.2013+Crew+Tonnage)
```



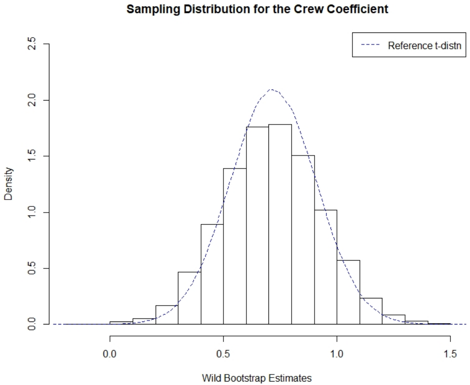
Example: Cruise Ship Properties

`lm(Passengers~Age.2013+Crew+Length+Tonnage)`



Example: Cruise Ship Properties

```
wild.boot(Passengers∼Age.2013+Crew+ Length+Tonnage,
B=10000)$bootEstParam
```



Wild Bootstrap 95% CI: (0.30, 1.15)

t_{153} 95% CI: (0.34, 1.09)

ROSE-HULMAN INSTITUTE OF TECHNOLOGY

What's coming...

- Easy construction of bootstrap confidence intervals and hypothesis tests for parameters
- Visualization of bootstrap sampling distributions
- Parallel capability
- Function to perform generalized bootstrap
- Vignette to guide new (or student) R users

References

Chatterjee, S. and Bose, A. (2000). "Variance Estimation in High Dimensional Linear Models." *Statistica Sinica*. Vol. 10, pp.497-515

Efron, B. (1979). "Bootstrap methods: Another look at the jackknife." *Annals of Statistics*. Vol. 7, pp.1-26.

Liu, R. Y. and Singh, K. (1992). "Efficiency and Robustness in Resampling." *Annals of Statistics*. Vol. 20, No. 1, pp.370-384.

Rubin, D. B. (1981). "The Bayesian Bootstrap." *Annals of Statistics*. Vol. 9, No. 1, pp.130-134.

Wu, C.F.J. (1986). "Jackknife, Bootstrap, and Other Resampling Methods in Regression Analysis." *Annals of Statistics*. Vol. 14, No. 4, pp.1261 - 1295.

Thank you!

Contact Info:

Megan Heyman (`heyman@rose-hulman.edu`)

Look for my R package in CRAN: `lmboot`