Exercises Lecture 09

Resampling 2: The Bootstrap

SCR team

14 November, 2019

Exercises part 1

When the parametric bootstrap would not be an optimal procedure... but still could be used.

1.1 Check the slides for the Parametric Bootstrap of Reaction Time.

We have created our own simulated reaction time data-set again, but now we use $n = 1000$ reaction times:

```
set.seed(160945)
alpha <- 3; beta <- 1 # true values
n <- 1000
X <- rgamma(n, alpha, beta) # data
```

The Gamma distribution with parameters shape $= \alpha$ and rate $= \beta$, and has density

$$f(x) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{(\alpha-1)} e^{-(\beta x)}$$

the mean and variance are $\mathrm{E}(X) = \alpha/\beta$ and $Var(X) = \alpha/\beta^2$.

a

Use the function `dgamma()` and code your own density function `my_dgamma()` that gives the same results as `dgamma()`. Note that very small differences between the functions (e.g. a difference of $1e-15$) are allowed.

b

You may have learned in your Probability & Statistics course(s) that the maximum likelihood estimates of the parameters of the gamma distribution for our specific data would be

```
beta_hat <- mean(X) / var(X)
alpha_hat <- mean(X) * beta_hat
c(shape = alpha_hat, rate = beta_hat)
```

```
##    shape     rate
## 3.037211 1.057207
```

In the slides of the lecture we have seen a parametric bootstrapping procedure to obtain an estimate of the standard error of the sampling distribution of the median. We used the maximum likelihood estimates of the shape and rate parameters to be able to draw samples from the gamma distribution (with the function `rgamma()`). What is your estimate of the expected median? What is your estimate of the standard error of the median?

c.

Create a histogram of the parametric bootstraped replicates of the median to visualize the estimate of the sampling distribution of the median. Add a vertical line of your observed median (in red), as well as the expected median (in blue) in your estimmate of the sampling distribution of the median.

d.

Perform a Monte-Carlo study, by sampling $B = 1000$ estimates of the median, but make sure to use the true shape (`alpha = 3`) and rate parameter (`beta = 1`).

d.i

Is the expectation of the Monte-Carlo medians the same as the expectation of the parametric bootstrapped medians? What is the difference (= estimate of the bias)?

d.ii

How about the standard error of the Monte-Carlo median vs. the parametric bootstrapped median? What would be your estimate of the bias?

d.iii

Visualize your Monte-Carlo replicates of the median in a histogram and add your observed median as a red vertical line and the expected median as a blue vertical line.

e

Without conducting your experiment in `R` code, what would be a good way to get an idea of how 'trustworthy' all these estimates are? What would happen to your estimates for lager n, and what would happen for larger B?

1.2 A Parametric Bootstrap for Regression Analysis

Load the data set `Advertising.csv` either from

http://www-bcf.usc.edu/~gareth/ISL/Advertising.csv

or from this `Rproject`'s directory (path is `./data/Advertising.csv`) into R and explore the data set a bit.

This dataset (see James, Witten, Hastie and Tibshirani, 2017) contains measurements of sales (in thousands of units), and of TV, radio and newspaper budgets (in thousands of dollars), for 200 different markets.

Let y_i be the sales of a particular market i, let x_{i1} be the budget expenditure on TV advertizements in market i, let x_{i2} be the radio budget expenditure in market i, and let x_{i3} be the budget expenditure on newspapers in market i.

In this exercises we are interested in the linear model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i,$$

where β_0 is an intercept, and β_1, β_2, β_3 are linear effects, and ϵ_i is an "error" term for which we have the assumption

$$\epsilon_i \sim N(0, \sigma)$$

a)

Use `lm()` to perform a linear regression analysis of `sales` on `TV`, `radio` and `newspaper` according to the linear model described above. Store the results from your linear regression analysis into a variable (aka object), and take a look at the summary (`summary()`) of your results. Is there a significant contribution of "advertisement in newspapers" on sales in this data set (controlled for TV and radio advertisements)?

b)

The sales that only contains the intercept and the linear effect of newspaper advertizements can be denoted (and defined) as follows:

$$\widehat{z}_i := \widehat{\beta}_0 + \widehat{\beta}_3 x_{i3} = y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + \widehat{\beta}_2 x_{i2} + \widehat{\beta}_3 x_{i3} + \widehat{\epsilon}_i),$$

Using the results that you have obtained in **a** create your own vector $\widehat{\mathbf{z}}$ that contains each $\widehat{z}_i$ for $i \in 1 \dots 200$.

c)

Use `all.equal()` to verify that the estimated standard deviation of the errors (misnomed as "residual standard error" in the output of `summary.lm()`) can be computed as

$$\widehat{\sigma} = \sqrt{\frac{\sum \widehat{\epsilon}_i^2}{n - p}}$$

where $n = 200$ markets, and $p = 4$ (the number of parameters: intercept + linear effects).

c)

Suppose you are a trainee at an advertisement company, and your boss wants to stop the advertising in the newspaper. To be sure, he would like you to validate the estimated standard error of the linear effect $\widehat{\beta}_3$ (= the contribution in sales the newspaper).

Do this as follows: Create $B = 1000$ parametric bootstrap samples of the $\mathbf{z}$, defined as

$$z_i^b = \widehat{z}_i + \epsilon_i^b$$

where ϵ_i^b is your own sampled residual from $N(0, \widehat{\sigma})$, the normal distribution with mean zero and a standard deviation equal to your estimate of the variance of the errors ($\hat{\sigma}^2$).

Then regress each $\mathbf{z}^b$ (for $b = 1 \dots B$) on the original `newspaper` variable ($\mathbf{x}_3$), and save the estimate of the linear effect (the coefficient) of advertizement expenditure for the newspaper.

What is the mean of these bootstrapped main effects, and what is the standard deviation? Are these results similar to those obtained in **a)**?.

Exercises part 2

2.1 Combining the Empirical and Parametric Bootstrap for Regression Analysis

Repeat exercise 1.2d, but instead of sampling the residuals from a normal distribution with variance equal to the observed estimated residual variance, apply the empirical bootstrap. Do the results remain similar?

2.2 Failing the Bootstrap: Regression Analysis with too small n

Suppose we have the regression model

```
fm1 <- lm(Employed ~ ., data = longley)
M1 <- model.matrix(fm1)
betas <- coefficients(fm1)
summary(fm1)
```

```
##
## Call:
## lm(formula = Employed ~ ., data = longley)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.41011 -0.15767 -0.02816  0.10155  0.45539
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -3.482e+03  8.904e+02  -3.911 0.003560 **
## GNP.deflator  1.506e-02  8.492e-02   0.177 0.863141
## GNP          -3.582e-02  3.349e-02  -1.070 0.312681
## Unemployed   -2.020e-02  4.884e-03  -4.136 0.002535 **
## Armed.Forces -1.033e-02  2.143e-03  -4.822 0.000944 ***
## Population   -5.110e-02  2.261e-01  -0.226 0.826212
## Year          1.829e+00  4.555e-01   4.016 0.003037 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3049 on 9 degrees of freedom
## Multiple R-squared:  0.9955, Adjusted R-squared:  0.9925
## F-statistic: 330.3 on 6 and 9 DF,  p-value: 4.984e-10
```

Program your own empirical bootstrap procedure on the residuals to see whether you can validate the standard errors of all coefficients (except the intercept) for the results in full model `fm1`.

2.3 Bootstrapping Quakes and the Spearman correlation coefficient

We will use the data set “quakes” in this exercise (type in the Console `?quakes` to learn more about these data). Somehow, we are interested in the distribution of the Spearman correlation coefficient between the variables longitude (long) and latitude (lat) for 1000 seismic events. To calculate the Spearman correlation coefficient use the function `cor()` with argument method = “spearman”.

Draw $B = 3000$ bootstrap samples of size n with replacement from the original data.

a)

In each of the 3000 bootstrap samples calculate the Spearman correlation coefficient and collect these coefficients in a vector (`CorSp_bs`).

b)

Make a histogram of the distribution of the vector created in **a)**, and add a vertical linear to indicate the observed Spearman correlation `obs_spcor`

c)

What would be the 95% confidence interval for the true Spearman correlation coefficient when using the percentiles only of the empirical bootstrapped values?

d)

What is the 95% confidence interval based on the normal approximation. For the standard error of the spearman correlation coefficient, use the the estimate of the standard error (= standard deviation) that you can obtain from your empirical bootstrap replicates of the spearman correlation coefficients.

e)

Also create the 95% confidence interval based on the approximate pivot function while using the empircal bootstrap procedure (slide 59 of the lectures).

f)

When comparing these intervals alltogether, would you conclude with approx. 95% confidence that there is a (very) small negative spearman correlation between longitude and latitude regarding seismographic events?

Remaing Exercies / Self-Study (Difficult!)

3.1

a

According to `R`, the closest estimate to the 'true' median of our median reaction time experiment of Exercises 1.1. would be:

```
alpha <- 3; beta <- 1 # true values
med_mc_approx <- qgamma(0.5, alpha, beta)
med_mc_approx
```

```
## [1] 2.67406
```

Come up with an estimate of the coverage for the three types of 95% confidence intervals for the observed median reaction time of Exercise 1.1 for as well the parametric bootstrap and the empirical bootstrap. What is the proportion taken over e.g. $B_{MC} = 1000$ confidence intervals for $B_{boot} = 1e3$ bootstrap replicates that each confidence interval envelopes the true parameter for the median, i.e. `med_mc_approx`?

The three types of confidence intervals are:

1. the 95% confidence interval using only the quantiles of the paramteric bootstrap replicates:

- `L(median_observed) = quantile(t_pboots, 0.025)`
- `U(median_observed) = quantile(t_pboots, 0.975)`

2. the 95% confidence interval using the normal approximation, which will be calculated as follows (L = lower bound; U = Upper bound):

- `L(median_observed) = median_observed - qnorm(0.975)*sd(t_pboots)`
- `U(median_observed) = median_observed - qnorm(0.025)*sd(t_pboots)`

3. the 95% confidence interval using the quantiles of the parametric bootstrap replicates and the pivot function with scale equal to 1, which is calculated as follows:

- `L(median_observed) = median_observed - (quantile(t_pboots - mean(t_pboots), 0.975))`
- `U(median_observed) = median_observed - (quantile(t_pboots - mean(t_pboots), 0.025))`

Note that for good estimates of the coverage of each confidence interval a much larger value for B_{MC} would be needed to reduce simulation error on the coverage estimate. Thus, we cannot form strong conclusions on our results! Also note that our definition of the coverage is also way to blunt (but workable).

b

Instead of using the empirical bootsrap, would you be able to repeat the whole experiment to estimate the coverage for each of these three interval types while using the parametric bootstrap?

3.2. Bootstrapping: a task from an Old Exam (SCR 2010)

a)

Consider a study with following factors:

- `sex`: factor with levels "male" and "female"

- `treat`: factor with levels "active" and "placebo"

- `age`: factor with levels "young" and "old"

- `bmi`: factor with levels "under", "normal", "over" and "obese"

Construct the `data.frame` that contains all possible combinations of the factors. *Hint: the dimension are 32 rows and 4 columns*

b)

For each possible combination of factor levels in (a) simulate 50 objects from the linear regression model

$$y_i \sim N(\mu_i, \sigma^2), \qquad i = 1, \dots, n = 50\text{Ö}32,$$

where $\sigma = 3$ and

$\mu_i = \beta_0 + \beta_1 \texttt{female}_i + \beta_2 \texttt{placebo}_i + \beta_3 \texttt{old}_i + \beta_4 \texttt{normal}_i + \beta_5 \texttt{over}_i + +\beta_6 \texttt{obese}_i$

with $\texttt{female}_i$ denoting the dummy variable for females, $\beta_2 \texttt{placebo}_i$ the dummy variable for placebo patients, $\beta_3 \texttt{old}_i$ the dummy variable for old patients, $\beta_4 \texttt{normal}_i$ the dummy variable for patients with normal BMI, $\beta_5 \texttt{over}_i$ the dummy variable for obese patients.

For the regression coefficients take the values $\beta_0 = 10, \beta_1 = 1, \beta_2 = -2, \beta_3 = -1.3, \beta_4 = 0.1, \beta_5 = -1, \beta_6 = -1.2$.

Hint: check the function `model.matrix` and it's use in predicting and fitting linear models

c)

Using the data you simulated in (b), fit the following four linear regression models:

- `M1` additive model: include only the main effects of `sex`, `treat`, `age`, and `bmi`.
- `M2` 2-way interaction model: include only the main effects of `sex`, `treat`, `age`, and `bmi`, and all the 2-way intereaction terms.
- `M3` 3-way interaction model: include only the main effects of `sex`, `treat`, `age`, and `bmi`, all the 2-way intereaction terms and all the 3-way intereaction terms.
- `M4` 4-way interaction model: include only the main effects of `sex`, `treat`, `age`, and `bmi`, all the 2-way intereaction terms, all the 3-way intereaction terms, and all the 4-way intereaction terms. The formula `y ~ (x1 + x2 + x3 + x4)^4` might be helpful.

Also, extract the design matrices for each model. Remember from linear regression: $\mathbf{b} = (\mathbf{X^T X})^{-1}\mathbf{X^T Y}$. The $\mathbf{X}$ is the design matrix.

Stronger hint: really, check the function `model.matrix()`

d)

We are interested in quantifying the predictive ability of these models. As a measure of predictive ability we wil use the adjusted `R^2` defined as

$$R^2_{adj} = 1 - (1 - R^2)\frac{n-1}{n-p-1}$$

where n denotes the sample size, p the total number of regressors in the linear model (but not counting the constant term / intercept), and

$$R^2 = COR(\hat{y}, y)^2,$$

the squared correlation between the predicted outcome $\hat{y}$ and the outcome variable y.

Use the Bootstrap to estimate a validated R^2 for each of the four models using 200 samples with replacement from the original sample. Is your valdiated R^2 closer to the observed R^2 or the observed R^2_{adj}? Do you see any differences when the number of parater P becomes larger?

It is often claimed in biostatistics and in the social sciences that R^2_{adj} is more valid to use for predictive ability than R^2. So, one would expect that the validated vlaues for R^2 are closer to those of the obseved R^2_{adj} when the number of parameter P becomes larger. Is this the case?